MATERIAL DIDÁCTICO
MATEMÁTICAS **6**

MÉTODOS DE ANÁLISIS DE DATOS: APUNTES

Zenaida Hernández Martín





MÉTODOS DE ANÁLISIS DE DATOS

(APUNTES)

MATERIAL DIDÁCTICO

Matemáticas

nº 6

Zenaida Hernández Martín

MÉTODOS DE ANÁLISIS DE DATOS

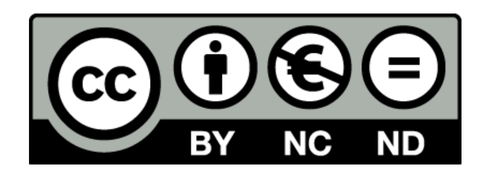
(APUNTES)

UNIVERSIDAD DE LA RIOJA

SERVICIO DE PUBLICACIONES

2012

Hernández Martín, Zenaida
Métodos de análisis de datos : apuntes / Zenaida Hernández Martín. -
Logroño : Universidad de La Rioja, Servicio de Publicaciones, 2012.
172 p. ; 29 cm. (Material Didáctico. Matemáticas ; 6)
ISBN 978-84-615-7579-4
1. Métodos estadísticos. I. Universidad de La Rioja. Servicio de Publicaciones, ed.
519.2



publicaciones.unirioja.es

E-mail: publicaciones@unirioja.es

ISBN 978-84-615-7579-4

Edita: Universidad de La Rioja, Servicio de Publicaciones

Prólogo

Este libro no pretende ser nada más que lo que es: unos apuntes completos de lo que se imparte en las clases de Métodos de Análisis de Datos. Un documento de ayuda a los estudiantes, para que puedan dedicarse a escuchar atentamente las explicaciones del profesor y a resolver los ejemplos y ejercicios planteados en clase y, a la vez, que sus apuntes estén completos, las definiciones correctas, las explicaciones estén recogidas y que las fórmulas estén correctamente escritas.

Aunque se incluyen algunos ejemplos, en estos apuntes no se incluye, como suele ser habitual, una lista de problemas, al menos en esta primera versión.

Por otra parte, los contenidos teóricos impartidos en clase se deben completar con unas prácticas en aula informática, en las que se aprenderá a hacer un análisis estadístico utilizando un *software* adecuado. Estas prácticas abarcan desde la obtención de datos a partir de las fuentes más habituales de información estadística, al análisis de los mismos, tanto de forma descriptiva como mediante la realización e interpretación correcta de los contrastes de hipótesis más habituales (los apuntes de estas prácticas tampoco están contenidos en este libro).

La mayoría de los gráficos y tablas que aparecen en estos apuntes se han realizado con el *software* de StatSoft, Inc. (2007). STATISTICA (Sistema informático de análisis de datos), versión 8.0. www.statsoft.com, que es el que se está utilizando actualmente en las clases prácticas.

Hablemos del contexto. La asignatura de Métodos de Análisis de Datos es una asignatura de Estadística Básica, que se imparte en varias titulaciones, sobre todo de las llamadas *de letras* y cuyos contenidos están pensados para familiarizar a los estudiantes con las técnicas más elementales de la Estadística, con su manejo y su interpretación.

El objetivo de la asignatura es que los estudiantes conozcan distintas medidas y técnicas estadísticas, sepan cuándo aplicarlas y sobre todo, cómo interpretarlas. No se pretenden grandes sesiones de cálculo y tampoco se hace mucho hincapié en el fundamento matemático, sino que se busca la comprensión de los estadísticos, cuándo, para qué y por qué aplicarlos.

Según los objetivos descritos para esta asignatura en los distintos Grados en los que se imparte, el estudiante debe adquirir una serie de competencias y habilidades, entre las que se encuentran las siguientes:

- Deberá ser capaz de enfrentarse a una situación y reconocer, si lo hay, un problema

estadístico. Por otra parte, a la vista de una serie de resultados estadísticos, debe ser capaz de interpretarlos, resumiendo la información y/o describiendo la situación de una forma coherente.

- Deberá adquirir conocimientos estadísticos básicos suficientes para comprender y defender o rechazar argumentos estadísticos de la vida cotidiana.
- Deberá conocer y aplicar las técnicas más utilizadas para la presentación y resumen de datos unidimensionales y bidimensionales, tanto cuantitativos como cualitativos.
- Deberá ser capaz de elaborar, presentar y defender un informe de la materia bien estructurado, utilizando el lenguaje correcto y la terminología adecuada.

Para conseguirlo, se ha pensado en un temario que incluye 10 temas y que son los que constituyen este curso.

Este documento está basado en el desarrollo del temario de la asignatura durante los cursos 2009-2010 y 2010-2011, de modo que se ajusta en tiempo y contenidos a los objetivos que se pretenden, por lo que es válido no solo para los estudiantes, sino también, como marco de referencia, para cualquier profesor que tenga que abordar por primera vez esta asignatura o alguna similar.

Como decía al principio, este libro no pretende ir más allá de los apuntes, completos, de clase. Tras el Índice, se comentan cuatro libros que se ajustan bastante al temario y al nivel de esta asignatura. Para acceder a otras explicaciones y/o ampliar conocimientos tienen en la Biblioteca de la Universidad bibliografía actualizada más que suficiente.

Por último, no sería justo terminar esta pequeña introducción sin agradecer a mis compañeros Montse San Martín, Juan Carlos Fillat y David Ortigosa, sus aportaciones y correcciones y sobre todo por su apoyo para que estos apuntes pudieran salir a la luz.

Logroño, julio de 2011

Índice

Bibliografía comentada

En la Biblioteca de la Universidad se dispone de abundante bibliografía actualizada, con la que se pueden completar estos apuntes y profundizar más en el temario.

Aquí se recomiendan algunos libros, que permiten completar la información de cada tema, tanto por su facilidad de comprensión como por ajustarse bastante a los contenidos que nos interesan.

- *Introducción a la Estadística Económica y Empresarial*, Martín-Pliego López, F. J.; Ed. Thomson. Madrid. 2004 (3ª edición).

  Incluye los temas: 2, 3, 8, 9 y 10.

- *Lecciones de Estadística Descriptiva. Curso teórico-práctico*, Tomeo Perucha, V. y Uña Juárez, I.; Ed. Thomson. Madrid. 2003.

  Incluye los temas: 2, 3, 8, 9 y 10.

- *Análisis de datos en Psicología I. Teoría y ejercicios.* Botella, J. y otros; Ed. Pirámide. Madrid. 2001.

  Incluye los temas: 2, 4, 5, 6, 8 y 9.

- *Estadística para las ciencias del comportamiento.* Pagano, R.; Ed. Thomson. Méjico. 1999 (5ª edición).

  Incluye los temas: 2, 4, 5, 6, 7, 8 y 9.

Tema 1

Estadísticas económicas y sociales

¿Para qué necesita un trabajador social o un economista la Estadística?

¿Entendemos las noticias de los periódicos?

¿Sabemos contestar a un argumento estadístico elemental?

El objetivo de este tema es variado.

- En primer lugar y como tema principal: comprender la utilidad de la Estadística en las ciencias sociales.
- En segundo lugar, debemos establecer algunas definiciones y conceptos elementales que nos permitan unificar el vocabulario y los criterios para comenzar a trabajar utilizando un correcto lenguaje estadístico.
- Por último, para hacer un estudio estadístico necesitamos datos. En algunos casos debemos obtenerlos nosotros, pero en otros muchos casos, ya hay mucha información elaborada por organismos oficiales. En este sentido comentaremos diversas fuentes de información estadística tanto de ámbito regional, como nacional e internacional.

1.1. La utilidad de la Estadística

A la hora de tomar decisiones en nuestro trabajo, e incluso en cualquier situación de nuestra vida cotidiana, nos encontramos con que esas decisiones las debemos tomar basándonos en una información que nos dan o que, de alguna forma, conocemos.

Aunque no nos demos cuenta, estamos manejando información estadística en situaciones tales como:

- El niño pide la paga y sus padres le preguntan: ¿y cuánto les dan a tus amigos sus padres?
- Nos cuestionamos las noticias ya que leemos o escuchamos que «seis de cada diez trabajadores en España son mileuristas» (en el comentario se especifica que del

total de los 27.94 millones de personas que perciben algún ingreso (asalariados, pensionistas, parados y autónomos), el 63 % tiene unos ingresos brutos mensuales inferiores a los 1100 euros). Mientras que por otro lado nos dicen que el sueldo medio mensual en España es de más de 1500 euros.

- Tenemos que renovar el alquiler con la subida del IPC.
- Nos dicen que los precios suben un 2 % (y no nos suben más el sueldo), pero a nosotros no nos llega para comprar lo mismo que el año pasado.
- Estamos viendo un partido de baloncesto y tenemos la información de la diferencia de puntos en cada minuto.

Son muchas las situaciones en las que vamos a tener que tomar decisiones importantes. Para ello tendremos que conocer, de alguna forma, la situación concreta que estamos analizando por lo que debemos manejar información sobre la misma.

Desgraciadamente no siempre podremos basar nuestras decisiones en la experiencia, pero cuando esto es posible, entra en juego la Estadística. Por lo tanto, las situaciones que nos interesan aquí son aquellas en las que vamos a manejar datos para ayudarnos a tomar nuestras decisiones.

Una vez que tenemos los datos, la investigación social se puede utilizar con dos enfoques: para describir el fenómeno o para tomar decisiones.

A partir de una masa de datos, la Estadística Descriptiva nos permite describir la situación analizada. Para ello se utilizan métodos de reducción de la masa de datos, cálculo de promedios, dispersión o tendencias, que nos permiten sacar conclusiones de estos datos.

Supongamos, por ejemplo, que conocemos las notas de selectividad de los 225 estudiantes que se matricularon en septiembre en una universidad pequeña. Esto constituiría una masa de datos.

Vamos a manejar una tabla ficticia para este ejemplo, pero más adelante veremos que en muchas ocasiones (no en todas) se pueden conseguir los datos reales sin mucha dificultad.

7.3	4.2	6.5	4.0	4.7	7.7	8.0	2.2	6.6	3.4	5.6	8.9	7.0	9.9	7.7
7.9	4.4	4.1	3.5	4.0	5.3	3.0	7.8	6.2	6.5	4.3	7.1	7.5	3.0	3.4
5.0	7.4	6.0	6.9	8.8	5.7	6.8	5.1	4.0	6.1	3.3	8.4	9.3	7.2	3.4
5.0	9.8	5.8	9.1	8.3	4.4	8.4	5.4	7.0	5.6	6.3	7.7	6.4	5.8	3.4
0.9	8.1	8.1	6.3	5.7	3.0	4.5	8.5	9.6	7.6	1.8	7.0	2.6	3.2	4.9
3.7	4.3	6.7	3.9	8.5	8.3	3.3	6.4	4.2	8.5	5.9	7.2	7.2	5.8	2.7
5.1	1.2	4.0	5.4	5.2	6.6	1.0	2.7	6.2	9.3	8.1	2.0	9.6	4.5	4.0
6.0	9.2	9.0	8.8	7.3	5.4	6.5	5.1	6.0	8.2	4.7	5.1	4.9	5.6	8.9
8.0	5.4	6.5	3.2	8.1	4.2	2.3	4.0	4.6	7.8	6.7	5.9	6.8	6.2	8.3
6.2	4.8	6.8	7.5	7.4	6.7	4.7	4.5	1.4	3.3	2.1	6.8	6.1	7.6	4.1
1.3	6.7	7.2	8.2	6.2	2.6	5.4	5.0	8.5	6.1	8.7	6.1	0.3	3.9	6.7
4.1	1.7	7.0	6.1	4.8	9.0	5.7	6.2	7.3	8.7	8.5	4.6	8.7	7.3	9.5
5.1	9.1	8.0	1.2	6.3	3.4	3.6	8.7	9.2	3.1	5.4	6.5	3.8	8.2	9.7
3.9	7.7	9.4	5.9	7.7	8.8	6.2	2.3	6.4	7.8	3.6	7.1	4.8	3.6	6.2
7.1	7.8	4.6	6.0	8.9	4.7	8.7	4.3	5.3	6.8	1.8	2.3	6.3	9.1	8.2

La simple observación directa de esta masa de datos (son números) no nos permite sacar conclusiones respecto a los mismos. Sin embargo, utilizando las técnicas de Estadística Descriptiva, incluso las más elementales, podemos describir el comportamiento de las calificaciones de los estudiantes con bastante precisión.

En los próximos temas veremos con detenimiento estas técnicas, pero ahora, como ejemplo, vamos a ver su utilidad:

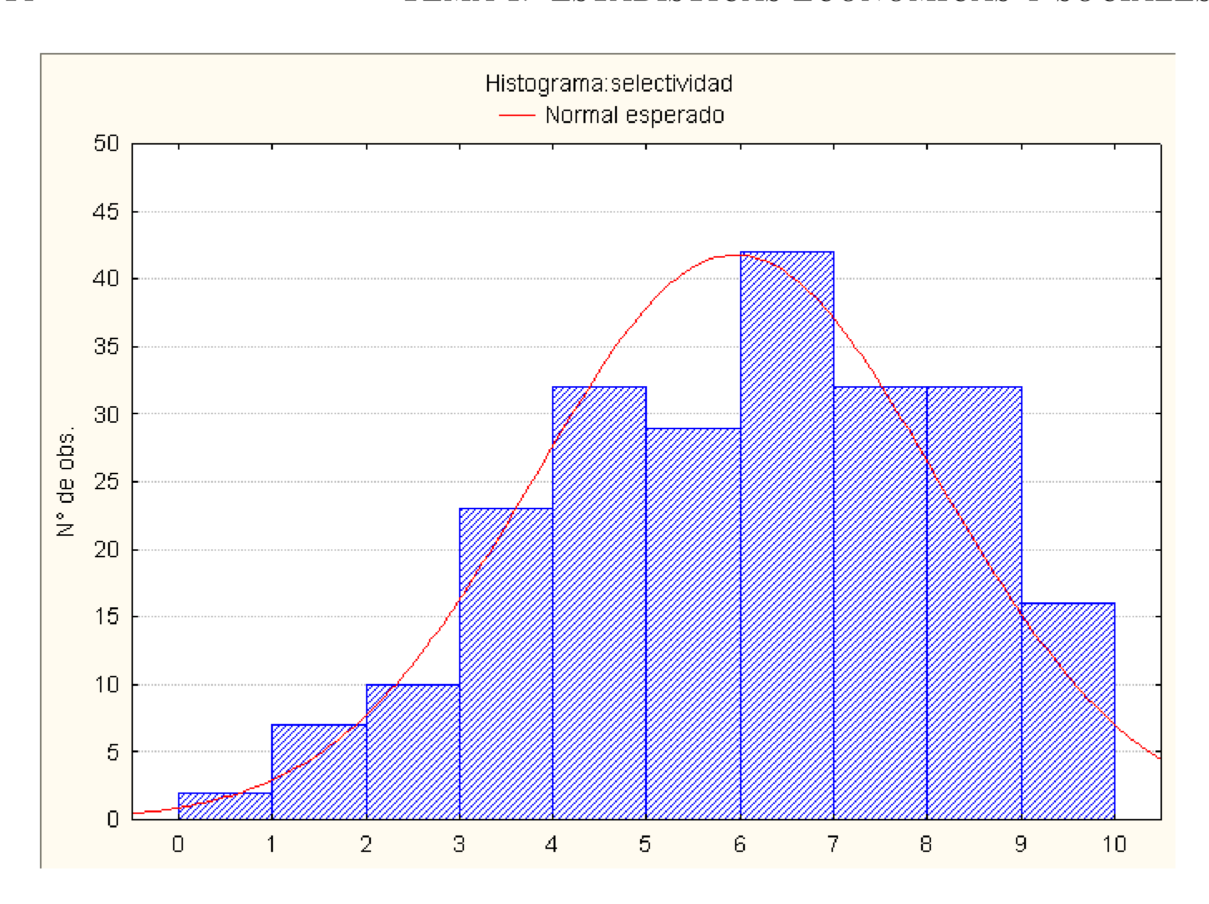
Un primer paso para sacar alguna conclusión de esta masa de datos consiste en **reducirla**. Para ello se procede a ordenarlos y agruparlos en categorías (este proceso se conoce como tabulación):

	Tabla de frecuencia: selectividad (Ejemplo_T1)			
de a	**Frecuencia**	Acumul. de Frecuencia	%	Acumul. de %
0,000000<=x<1,000000	2	2	0,88889	0,8889
1,000000<=x<2,000000	7	9	3,11111	4,0000
2,000000<=x<3,000000	10	19	4,44444	8,4444
3,000000<=x<4,000000	23	42	10,22222	18,6667
4,000000<=x<5,000000	32	74	14,22222	32,8889
5,000000<=x<6,000000	29	103	12,88889	45,7778
6,000000<=x<7,000000	42	145	18,66667	64,4444
7,000000<=x<8,000000	32	177	14,22222	78,6667
8,000000<=x<9,000000	32	209	14,22222	92,8889
9,000000<=x<10,00000	16	225	7,11111	100,0000
10,00000<=x<11,00000	0	225	0,00000	100,0000
Faltante	0	225	0,00000	100,0000

De esta forma, podemos comenzar a hacernos una idea de la distribución de la variable estudiada (calificaciones).

La simple observación de la tabla nos permite decir que la mayoría de las calificaciones están en la parte central y que en los extremos hay pocas observaciones; que hay 74 suspensos que corresponden al 32.89 % de las observaciones o que lo más habitual es tener una nota en el intervalo [6,7).

En la mayoría de las ocasiones, también es interesante representar gráficamente esta información ya que la interpretación suele ser más sencilla:



En el histograma anterior, podemos visualizar algunas de las observaciones que ya hemos hecho, como destacar el hecho de que son muy pocas las calificaciones por debajo de 3 o por encima de 9; que la mayoría de las calificaciones se encuentran entre 4 y 9 y que se reparten de forma bastante «uniforme» salvo en el intervalo [6,7) en el que hay un mayor número de calificaciones.

Aunque todavía no hemos comenzado con la asignatura propiamente, un procedimiento estadístico con el que todos estamos familiarizados, es preguntarnos por la calificación promedio. La media aritmética se obtiene sumando todas las calificaciones y dividiendo por el número total de estudiantes y nos permite hacernos una idea de la tendencia en el grupo. Nos da un valor alrededor del cual se encuentran todas las demás calificaciones. En este caso la calificación media es de 5.92 lo que nos dice que la calificación media de selectividad ha sido bastante baja.

Así, con estos sencillos recursos estadísticos (tablas de distribución de frecuencias, gráficos y la media aritmética) hemos podido detectar y describir algunos aspectos del comportamiento de las calificaciones, que la mera observación de la masa de datos no nos permite.

Las técnicas de Estadística Descriptiva nos van a permitir reducir la masa de datos a unos pocos indicadores con los que podremos describir adecuadamente el comportamiento de la variable.

La Estadística también se utiliza para contrastar hipótesis. Constantemente hacemos hipótesis o conjeturas sobre ciertas situaciones, pero cuando nuestras decisiones las tenemos que basar en estas hipótesis, es convenientes contrastarlas.

Podemos afirmar (porque es una creencia o porque nos da la impresión) que fuman más los hombres que las mujeres o que una determinada enfermedad tiene mayor incidencia en una provincia o en otra. Sin embargo, si tenemos que tomar una decisión basada en estos hechos, deberíamos saber cuál es el nivel de confianza de estas afirmaciones, hasta qué punto podemos apoyarnos en ellas. Esto lo haremos con los contrastes de hipótesis.

Está claro que si tenemos información completa de un fenómeno, no hay nada que contrastar. Si en la Universidad de La Rioja se han matriculado 3000 hombres y 3600 mujeres, podemos afirmar con certeza absoluta que hay más mujeres que hombres.

Sin embargo, hay informaciones que nos interesa contrastar ya que depender del sentido común o de las experiencias cotidianas tiene sus limitaciones y son muchas las ocasiones en las que las «creencias populares» no coinciden con la realidad.

Por ejemplo: «En la Universidad de La Rioja es más probable que tengan el carné de conducir los hombres que las mujeres»; esta afirmación se puede deber a una impresión por lo que se ve, pero no se sabe a ciencia cierta. Si queremos utilizarla con propiedad habrá que contrastarla.

Por otra parte, en la Universidad de La Rioja hay unos 6600 alumnos, por lo que quizás no nos sea posible entrevistarlos a todos para saber si tienen carné o no y distinguir por sexo a los conductores. En este caso habría que tomar una porción o muestra del grupo grande que queremos analizar (población), probaremos la hipótesis para la muestra y decidiremos si es posible y correcto extender el resultado a la población de la que se obtuvo la muestra.

El problema de generalizar, o hacer inferencia, es que al tomar una muestra estamos asumiendo que existe un error inevitable, por muy bueno y correcto que haya sido el muestreo. No podemos garantizar nuestra respuesta con una seguridad del 100 %. Si en la tabla de las notas tomamos varias muestras de 5 calificaciones, veremos que las medias para cada una de las muestras son distintas a la media global: 5.92. Esto es lo que llamaremos error de muestreo.

Supongamos para simplificar que la mitad de los alumnos son hombres y la otra mitad mujeres, y que a partir de las listas de alumnos tomamos una muestra de 100 hombres (1 de cada 33) y otra de 100 mujeres (1 de cada 33), ahora les preguntamos si tienen el carné o no.

Consideramos las siguientes tres respuestas:

	Respuesta 1		Respuesta 2		Respuesta 3	
	H	M	H	M	H	M
Carné SI	60	40	55	45	51	49
Carné NO	40	60	45	55	49	51
Totales	100	100	100	100	100	100

Las 3 respuestas están de acuerdo con nuestra hipótesis, pero lo que a nosotros nos interesa es saber si estas diferencias son lo suficientemente importantes como para generalizarlas a todos los estudiantes. Es decir, nos preguntamos si las diferencias encontradas

se deben al comportamiento de toda la población o solo se deben a la muestra elegida.

¿Hasta qué punto estamos dispuestos a aceptar la hipótesis con estos resultados?, ¿en qué punto es suficientemente grande la diferencia como para considerarla real?, la **Inferencia Estadística** nos permitirá tomar nuestra decisión de una forma sencilla y con un nivel de confianza determinado.

1.2. Definiciones iniciales

En los comentarios anteriores hemos estado utilizando algunos términos estadísticos sin conocer cuál es su definición correcta. Para poder trabajar en Estadística es conveniente tener claros los conceptos y utilizar un lenguaje común, que no dé lugar a confusión, por lo que vamos a proceder a dar algunas definiciones básicas.

Para poder realizar cualquier análisis estadístico debemos disponer de unos datos. Y estos datos corresponden a los valores obtenidos al estudiar determinadas características en los elementos de un conjunto de entes.

Para fijar el lenguaje que utilizaremos, estableceremos los siguientes términos:

Población es el conjunto de entes (personas, animales o cosas) sobre los que se va a llevar a cabo la investigación estadística.

Elemento es cada uno de los componentes de la población (pueden ser simples o compuestos).

Tamaño de la población es el número de elementos que la componen.

Caracteres son las cualidades o rasgos comunes a toda la población que vamos a estudiar. Pueden ser cuantitativos (**variables**) o cualitativos (**atributos**).

Aunque existe el análisis estadístico de los caracteres cualitativos (se verá al final del temario), cuando se habla de análisis estadístico, generalmente nos referimos al análisis de las características cuantitativas observadas en los elementos de una población.

Por lo tanto, generalmente trabajaremos con variables estadísticas que, atendiendo a los valores que pueden tomar, pueden ser **discretas o continuas**; y esta diferencia hace que en muchas ocasiones tengan un tratamiento diferente.

- Diremos que una variable estadística es **discreta** si dados dos valores distintos de la variable, entre ellos no puede haber más que un número finito de valores de la variable, por muy alejados que estén entre sí. Por ejemplo: número de hijos.
- Diremos que una variable estadística es **continua** si, dados dos valores distintos de la variable, entre ellos hay infinitos posibles valores de la variable, por muy próximos que estén entre sí. Por ejemplo: peso, tiempo...

Por otra parte, dentro de los atributos (también llamados variables cualitativas), cabe distinguir dos categorías: los atributos que son simples nombres y/o categorías (**atributos categóricos**) y los **atributos ordinales** que además permiten algún tipo de ordenación.

Por ejemplo, el estado civil es un atributo categórico, mientras que el grado de satisfacción o el nivel de estudios son atributos ordinales.

Es muy importante, en el caso de los atributos, no confundir los números que se pueden utilizar para codificar las distintas categorías con valores resultantes de una medición. NO podremos realizar operaciones aritméticas con estos números.

Otra cuestión muy importante, que se debe tener en cuenta antes de realizar un análisis estadístico es qué es lo que queremos o podemos hacer, en función del **tamaño de la población** objeto de estudio.

Si la población es pequeña y podemos obtener datos de todos los elementos de la misma, lo que haremos será un análisis descriptivo (**Estadística Descriptiva**).

Pero, si la población es muy grande (infinita o tan grande que no podemos abordarla en su totalidad), no nos queda más remedio que tomar una «muestra representativa», analizar dicha muestra y luego estudiar bajo qué condiciones podemos extender los resultados obtenidos con la muestra a toda la población o si podemos inferir algún resultado para la población. En esto consiste la **Inferencia Estadística**.

Una vez que tenemos claros estos conceptos, para realizar un análisis estadístico, generalmente seguiremos los siguientes pasos:

Paso 1: Establecemos la población que queremos estudiar.

Paso 2: Determinamos las características que nos interesa analizar de dicha población.

Paso 3: Recogemos los datos.

Paso 4: Realizamos el análisis de datos.

Paso 5: Exponemos nuestras conclusiones.

1.3. Fuentes de información estadística

Como ya hemos dicho, para realizar un análisis estadístico necesitamos manejar una masa de datos.

Estos datos los podemos haber recogido nosotros personalmente mediante estudios directos de la población o de una muestra representativa de la misma, pero en muchas ocasiones tendremos que recurrir a datos ya elaborados.

Los organismos oficiales tienen departamentos de Estadística dedicados a la recolección de datos que utilizan para elaborar sus informes correspondientes.

En la mayoría de las ocasiones estos datos nos los presentan semi-tratados y solo en algunos casos (afortunadamente cada vez más) tenemos acceso a los microdatos, es decir

a los datos originales de la encuesta.

No vamos a ver aquí una lista exhaustiva de fuentes de información estadística, sino más bien una idea de los lugares a los que podemos acudir.

En España, lo primero que se nos ocurre es acudir al Instituto Nacional de Estadística (INE): http://www.ine.es, un pequeño paseo por esta web nos permite acceder a gran cantidad de información estadística.

Además, si en la pestaña de ayuda, seleccionamos «Enlaces», accederemos a una página en la que se encuentran las direcciones actualizadas de las principales fuentes estadísticas, tanto nacionales como internacionales:



Si lo que necesitamos son los microdatos, la página del INE también nos permite acceder a muchos de ellos. En la pestaña de Productos y Servicios /Información podemos seleccionar ficheros de microdatos.

Por otra parte, si lo que necesitamos son los microdatos de las estadísticas que ofrecen otras fuentes, debemos acceder mediante su página web (si es que los datos son accesibles al público en general) o solicitarlos al organismo correspondiente, que valorará nuestra solicitud y puede que nos los ceda o no.

Por ejemplo, para obtener unos datos que son públicos en el CIS, nos piden que nos identifiquemos:

http://www.cis.es/cis/opencms/ES/formulario.jsp?dwld=/Microdatos/MD2811.zip

Evidentemente estas no son todas las fuentes estadísticas ya que nos hemos dirigido solo a organismos oficiales. Hay otras muchas organizaciones que también elaboran sus propias estadísticas y que nos pueden facilitar sus datos, aunque es recomendable utilizar los datos oficiales siempre que sea posible.

También se puede obtener información estadística en los Anuarios Estadísticos y otras publicaciones en papel, ya sean de organismos oficiales u otras organizaciones, que se encuentran en la Biblioteca o en las sedes de los mismos.

Tema 2

Estadística Descriptiva unidimensional

En este tema veremos cómo realizar el análisis descriptivo completo de una variable unidimensional.

Primero nos haremos una idea de su comportamiento con el resumen de los datos y algunos gráficos elementales y, a continuación, veremos cómo calcular las principales medidas que nos permitirán describir con precisión el comportamiento de dicha variable. Esta descripción la haremos interpretando correctamente todos los resultados obtenidos.

Los epígrafes del tema son los siguientes:

- Escalas de medición.
- Resumen de los datos: tablas de frecuencias.
- Gráficos unidimensionales.
- Medidas de tendencia central y de posición: media aritmética, mediana, moda y percentiles.
- Medidas de dispersión absolutas y relativas: Recorrido, varianza, desviación típica, cuasivarianza, cuasidesviación típica, recorrido relativo y coeficiente de variación.
- Medidas de forma: asimetría y curtosis.
- Medidas de concentración.

2.1. Escalas de medición

Aunque ya lo comentamos en el tema anterior, vamos a dejar un poco más claro el concepto de escala de medición ya que el tipo de escala influirá en el posible tratamiento posterior de la variable.

Escala nominal: las observaciones de un carácter vienen dadas en escala nominal cuando se pueden clasificar en varias categorías, excluyentes entre sí, entre las que no es posible establecer ninguna relación de orden y tampoco es posible operar matemáticamente.

En este tipo de escala vienen dados los **atributos categóricos**: sexo, estado civil, tipo de contrato laboral, lugar de nacimiento, sector de actividad económica,...

Escala ordinal: las observaciones de un carácter vienen dadas en escala ordinal cuando se pueden clasificar en varias categorías, excluyentes entre sí, entre las que es posible establecer alguna relación de orden, aunque no es posible operar matemáticamente. Las categorías se pueden ordenar pero no se puede cuantificar la diferencia entre ellas.

En este tipo de escala vienen dados los **atributos ordinales**: nivel de ingresos, nivel de estudios, grado de satisfacción,...

Escala de intervalo: las observaciones de un carácter vienen dadas en escala de intervalo cuando existe una unidad de medida que nos permite cuantificar la distancia existente entre dos observaciones, pero el cero es arbitrario.

En este tipo de escala vienen dadas **algunas variables** (caracteres cuantitativos): la temperatura en grados Celsius o en grados Fahrenheit, **fechas**,...

Si observamos dos temperaturas: 30º C y 15º C (equivalentemente: 86 ºF y 59 ºF respectivamente – ºF=32+1.8 ºC –), podemos establecer distancias (15 grados de diferencia en la escala Celsius, o 27 grados en la escala Fahrenheit), pero no podemos afirmar que 30 ºC es una temperatura doble que 15 ºC, la temperatura es la misma sea cual sea la escala de medida y aunque 30 es el doble de 15, sin embargo 86 no es el doble de 59.

Escala de razón (proporción o cociente): las observaciones de un carácter vienen dadas en escala de razón cuando existe una unidad de medida que nos permite cuantificar la distancia existente entre dos observaciones y además existe un cero absoluto.

En este tipo de escala vienen dadas **la mayoría de las variables** (caracteres cuantitativos): edad, peso, salario, nivel de inventario,...

En los dos últimos casos (escala de intervalo y escala de razón), al existir una unidad de medida, se puede operar matemáticamente con los datos y obtener una serie de medidas o indicadores que nos van a permitir describir el comportamiento, para nuestra población, de la característica que estamos analizando.

En algunos libros, solo se consideran 3 tipos de escala, ya que las escalas de intervalo y de razón se unen en una única categoría llamada **escala cuantitativa**.

2.2. Resumen de los datos: tablas de frecuencias

Una vez que hemos determinado cuál es la población que queremos estudiar y qué características queremos analizar, el siguiente paso es la recogida de datos.

Para cada individuo, obtendremos tantos valores como características estemos analizando. Así, si en una población solo nos interesa la edad, para cada individuo tendremos un único valor: su edad; pero si nos interesa la edad, antigüedad en la empresa, estado civil y salario, para cada individuo tendremos 4 valores.

En el primer caso diremos que obtenemos una variable unidimensional (E=edad) y en el segundo caso, tenemos una variable de dimensión 4 (E, A, C, S).

En realidad, la variable de dimensión 4, está formada por 4 variables unidimensionales, con la particularidad de que los valores de cada 4-tupla, corresponden al mismo individuo o elemento de la población. En este curso, solo nos vamos a ocupar de los análisis de variables unidimensionales y bidimensionales.

Para comenzar nos vamos a referir al estudio de un único carácter poblacional y por lo tanto a una variable unidimensional (por ahora no vamos a distinguir entre variables cualitativas y cuantitativas).

Las variables, en general, se suelen nombrar con una letra mayúscula $(E, A, X, Y, \dots)$. Cuando observamos una variable en una población, obtenemos una serie de valores distintos para esa variable: 18, 19, 20,..., o soltero, casado, viudo,... Los distintos valores observados de la variable se suelen nombrar con la misma letra que la variable pero en minúscula.

Al observar una característica, X, de la población podemos obtener unos valores (distintos entre sí): $x_1, x_2, ..., x_k$. Además, cada uno de los valores distintos observados de la variable, puede aparecer una o más veces.

Definimos:

Frecuencia absoluta de un determinado valor, x_i, de la variable (y la representaremos por n_i): es el número de veces que se presenta ese determinado valor x_i.

Frecuencia relativa de un determinado valor, x_i, de la variable (y la representaremos por f_i): es la proporción de veces que aparece ese valor en el conjunto de observaciones y se calcula como el cociente de su frecuencia absoluta (n_i) y el número total de datos, N.

Frecuencia absoluta acumulada de un determinado valor, x_i, de la variable (y la representaremos por N_i): es la suma de las frecuencias absolutas de todos los valores de la variable menores o iguales que dicho valor x_i.

$N_i = \sum_{j=1}^{i} n_j = n_1 + \cdots + n_i$, N_k=N.

Frecuencia relativa acumulada de un determinado valor, x_i, de la variable (y la representaremos por F_i): es la suma de las frecuencias relativas de todos los valores de la variable menores o iguales que dicho valor, x_i.

$F_i = \sum_{j=1}^{i} f_j = f_1 + \cdots + f_i = \frac{N_i}{N}$, F_k=1.

Las frecuencias acumuladas solo tienen sentido si la escala es ordinal o cuantitativa.

Cuando en un conjunto de valores observados de una variable se realizan las operaciones de **ordenación** y **agrupación de los valores que se repiten** (determinación de la frecuencia de cada valor), se obtiene una **tabla estadística de distribución de frecuencias**.

A dicho conjunto de operaciones se le denomina tabulación.

Nota: en el caso de los atributos, los valores se pueden escribir en cualquier orden, pero si son atributos ordinales, el construir la tabla con los valores ordenados, facilita la comprensión de la misma.

Para ver cómo se emplean estos conceptos, consideremos el siguiente ejemplo:

Población: La plantilla de una pequeña empresa, formada por 20 jóvenes.

Variable: edad, expresada en años.

Valores observados:

18	20	22	19	18
20	18	19	21	20
20	21	18	20	21
19	20	21	18	20

Entonces, si llamamos X a la variable edad, los valores x_i distintos que hemos observado son: 18, 19, 20, 21 y 22.

La correspondiente tabla de frecuencias será:

x_i	n_i	N_i	f_i	F_i
18	5	5	5/20	5/20
19	3	8	3/20	8/20
20	7	15	7/20	15/20
21	4	19	4/20	19/20
22	1	20	1/20	1
	20		1	

Ejemplo: Para las notas de 100 alumnos, vamos a construir la tabla de frecuencias:

4	1	5	6	3	5	2	4	4	6
3	4	0	4	7	7	3	4	8	6
8	3	4	5	3	6	9	6	1	5
1	0	1	2	1	3	2	7	5	6
5	4	3	5	5	4	7	5	2	1
2	1	2	3	1	3	5	2	5	5
7	5	3	5	4	6	6	4	7	7
6	0	2	4	2	4	7	3	3	2
8	4	6	6	4	5	10	6	4	7
8	2	4	6	4	4	4	2	6	7

La correspondiente tabla de frecuencias será:

x_i	n_i	N_i	f_i	F_i
0	3	3	0.03	0.03
1	8	11	0.08	0.11
2	12	23	0.12	0.23
3	12	35	0.12	0.35
4	20	55	0.20	0.55
5	15	70	0.15	0.70
6	14	84	0.14	0.84
7	10	94	0.10	0.94
8	4	98	0.04	0.98
9	1	99	0.01	0.99
10	1	100	0.01	1.00
	100		1	

Las distribuciones de frecuencias se pueden clasificar de acuerdo con el número de los valores observados de la variable, así como con el número de observaciones totales:

- En el caso de pocas observaciones y pocos valores de la variable, no es necesario realizar la operación de agrupamiento. En este caso, únicamente procede realizar una ordenación de los valores de la variable. Esta ordenación no supone tratamiento estadístico alguno, pues para que exista «tratamiento estadístico» se debe disponer de una «masa de datos».

  Ejemplo: Variable: peso físico

  Valores observados: 50, 64, 80, 72.

  Número de observaciones : 4

  El único tratamiento posible es la ordenación de los valores.

- Lógicamente cuando hay muchos valores, no puede haber pocas observaciones.

- Cuando disponemos de muchas observaciones correspondientes a pocos valores de la variable aparecen las tablas de frecuencias. Generalmente este tipo de distribuciones se presenta para variables discretas, ya que es poco realista que al realizar muchas observaciones de una variable continua, se obtengan pocos valores diferentes.

  Como ejemplo de una distribución de este tipo, puede servir el ejemplo anterior de los 20 empleados. O bien, estudiar la variable edad, con precisión de años, en los alumnos de primera matrícula en la Universidad de La Rioja.

- Si son muchas las observaciones y muchos los valores observados de la variable, *en ocasiones* se procede previamente a la agrupación de los valores de la variable en intervalos. Cuando esto ocurre, se habla de **tablas agrupadas en intervalos**. Este tipo de tablas es aplicable tanto a las variables discretas (cuando es muy elevado el número de valores), como a las variables continuas.

En este último caso, lo primero que se hace es agrupar los valores de la variable en intervalos, que pueden ser de amplitud constante o no, y calcular las frecuencias en cada intervalo.

Para agrupar los datos en intervalos o clases, debemos comenzar determinando el recorrido o rango de la variable, que se define como la diferencia entre el mayor y el menor valor de la variable:

$$\text{Re} = \text{máx}\, x_i - \text{mín}\, x_i$$

Este recorrido se divide entonces en intervalos. Lo más cómodo para el tratamiento posterior de la distribución es que los intervalos sean de amplitud constante, pues entonces: Re= número de intervalos × amplitud, lo cual permite deducir:

- el número de intervalos, si fijamos la amplitud

- la amplitud, si fijamos el número de intervalos.

No existen reglas fijas para determinar el número idóneo de intervalos, hasta el punto de que a veces se hacen varias pruebas hasta conseguir resaltar las características del fenómeno. Cuando no existen otras indicaciones, un valor comúnmente aceptado es un número próximo a raíz cuadrada de N (siendo N el número total de observaciones).

Cada intervalo queda especificado por sus límites. En general para el intervalo i-ésimo, estos límites se representan por l_{i-1} y l_i , donde l_{i-1} es el límite inferior y l_i el límite superior.

Un problema que puede surgir es que el valor de la variable coincida exactamente con el límite del intervalo. Para evitar que aparezcan situaciones conflictivas, es conveniente especificar el tipo de intervalo. Generalmente se utiliza abierto por la izquierda y cerrado por la derecha: $(a, b]$ o $]a, b]$. Lo cual significa que dentro del intervalo se incluyen los valores comprendidos entre a y b, incluido b y excluido a.

Para facilitar el manejo matemático de los intervalos, es preciso considerar un valor concreto de la variable como representante de cada intervalo. Generalmente se toma como tal el valor central del intervalo, y se le denomina **marca de clase**.

Ejemplo: en una escuela, las notas de Física de 100 estudiantes fueron:

4.4	1.1	4.6	5.8	2.5	4.8	1.8	4.1	3.5	5.9
2.9	3.5	0.2	3.7	6.8	7.0	3.1	4.4	8.4	6.4
8.2	2.6	4.2	5.1	2.9	5.9	9.2	5.6	0.5	5.2
0.8	0.1	1.2	4.7	2.1	0.6	3.2	1.5	6.7	6.1
4.7	4.3	3.3	4.8	4.7	4.3	6.9	4.9	2.1	0.9
1.5	1.1	2.2	2.9	1.4	3.1	4.6	1.9	4.9	5.1
7.1	5.2	3.2	5.1	4.4	5.7	6.0	4.3	6.5	7.3
6.2	0.3	1.7	3.9	2.2	4.0	6.5	3.0	3.1	1.6
8.0	4.1	5.9	6.0	4.1	5.1	1.0	6.3	4.1	7.4
8.1	2.0	3.6	5.9	3.8	4.0	4.3	1.8	6.0	7.1

Puesto que las puntuaciones pueden ir de 0 a 10, es cómodo el hacer 10 intervalos de

longitud constante igual a 1 punto. Los intervalos, las marcas de clase y los distintos tipos de frecuencias son los siguientes:

$(l_{i-1}, l_i]$	x_i	n_i	N_i	f_i	F_i
$[0, 1]$	0.5	8	8	0.08	0.08
$(1, 2]$	1.5	12	20	0.12	0.20
$(2, 3]$	2.5	10	30	0.10	0.30
$(3, 4]$	3.5	14	44	0.14	0.44
$(4, 5]$	4.5	21	65	0.21	0.65
$(5, 6]$	5.5	16	81	0.16	0.81
$(6, 7]$	6.5	10	91	0.10	0.91
$(7, 8]$	7.5	5	96	0.05	0.96
$(8, 9]$	8.5	3	99	0.03	0.99
$(9, 10]$	9.5	1	100	0.01	1

En las distribuciones agrupadas en intervalos se puede presentar el problema de que el último intervalo sea abierto, es decir, que no tenga límite superior (idéntico problema se puede presentar con el primer intervalo y el extremo inferior).

Por ejemplo, vamos a considerar la siguiente distribución de frecuencias de los ingresos mensuales de 1.000 familias:

Intervalo (en euros)	Marca de clase x_i	Frecuencia absoluta n_i
0-1000	**500**	100
1000-2000	1500	300
2000-3000	2500	400
3000-5000	4000	150
más de 5000	¿?	50
		1000

En la distribución anterior no se puede determinar directamente la marca de clase correspondiente al último intervalo.

Cuando *se conocen los valores individuales* de la distribución, lo que suele hacerse es tomar como marca de clase del último intervalo el promedio de todos los valores que corresponden al mismo (en este ejemplo se calcularía el promedio correspondiente a las 50 familias que cuentan con ingresos superiores a 5000 euros mensuales).

Si no se dispone de esa información individual, no existen criterios objetivos que nos permitan determinar la marca de clase.

Sin embargo, para el cálculo de ciertas características de la distribución, como veremos más adelante, es necesario conocer todas las marcas de clase. En ese caso, se toma como marca de clase de un intervalo abierto, el valor que, a *juicio del que realiza la investigación*, mejor representa el intervalo. Como se puede observar esta es una determinación arbitraria, pero en ciertos casos, no existe otra solución.

Como ya se ha indicado, se suele tomar como marca de clase el valor central del intervalo, ya que en principio se considera como el valor más representativo del mismo. *Pero en algunas ocasiones se observa que este criterio es totalmente inaceptable.* Así, en el ejemplo anterior, no parece razonable que la marca de clase 500 euros, sea un buen representante de las 100 familias con ingresos comprendidos entre 0 y 1000 euros. Lógicamente, *cabe suponer que la mayor parte de estas familias se acercarán más a los 1000 euros que a los 0 euros.* Para conseguir que la marca de clase sea representativa, debe adoptarse una solución similar a la adoptada en los intervalos abiertos.

2.3. Lectura de las tablas de frecuencias

Puesto que, como hemos comentado, la reducción de datos se realiza para hacer más manejable y comprensible la masa de datos, vamos a ver ahora cómo extraer información de una tabla de frecuencias, y cómo expresar dicha información, dependiendo de la forma de la tabla y de nuestras necesidades.

Supongamos que tenemos la siguiente tabla con la información sobre los estudiantes de cierta universidad, por sexo:

Sexo s_i	Frecuencia n_i
Hombre	25704
Mujer	24696
Total	50400

Una simple mirada a la tabla nos permite decir que en esa universidad hay más hombres que mujeres, y que en total hay 50400 estudiantes.

Podríamos ampliar esta información si completamos la tabla con las **frecuencias relativas y porcentajes**:

Sexo s_i	Frecuencia n_i	Frecuencia relativa f_i	Porcentaje $100 \times f_i$
Hombre	25704	0.51	51
Mujer	24696	0.49	49
Total	50400	1	100

Ahora esa diferencia la podemos cuantificar: el 51 % de los estudiantes de esa universidad son hombres, mientras que el 49 % restante son mujeres.

Por otra parte, nos puede interesar comparar a los estudiantes por sexo pero no en toda la universidad sino para los estudiantes de Ciencias y de Letras, es decir, nos podríamos preguntar si también es cierto que hay más hombres que mujeres tanto en las facultades de Ciencias como en las de Letras.

Para ello deberíamos tener información del sexo y tipo de facultad de cada uno de los estudiantes (variable bidimensional), y esta información la podríamos recoger en una tabla del tipo:

Facultad / Sexo	Ciencias	Letras	Total
Hombre	13608	12096	25704
Mujer	16632	8064	24696
Total	30240	20160	50400

Entonces, podemos decir que aunque en las facultades de Letras sí que es cierto que el número de hombres es mayor que el número de mujeres, esto no es cierto en las facultades de Ciencias, donde es mayor el número de mujeres que el de hombres.

También de esta tabla podemos obtener otra información y es que en esta universidad hay más estudiantes de Ciencias que de Letras.

Cuando los grupos tienen distinto tamaño, para hacer las comparaciones entre los grupos y hacernos una idea clara de las diferencias, es conveniente «estandarizar» las distribuciones por tamaño, para ello se suelen utilizar las **proporciones o los porcentajes**.

Recordemos que las proporciones comparan el tamaño de una categoría dada con el valor de toda la distribución (son las frecuencias relativas). Sin embargo hay mucha gente que prefiere indicar el tamaño relativo en forma de porcentaje, o lo que es lo mismo, la frecuencia de una determinada categoría por cada 100 casos.

Veámoslo sobre el ejemplo anterior. Comparamos las distribuciones por columnas.

Facultad / Sexo	Ciencias	Letras	Total
Hombre	13608 (45 %)	12096 (60 %)	25704 (51 %)
Mujer	16632 (55 %)	8064 (40 %)	24696 (49 %)
Total	30240 (100 %)	20160 (100 %)	50400 (100 %)

Ahora podemos saber que en las facultades de Ciencias, de cada 100 alumnos matriculados, 45 son hombres y 55 son mujeres, mientras que en las facultades de Letras, hay un 60 % de estudiantes hombres y solo un 40 % de mujeres.

Como podemos ver, se pueden apreciar mejor las diferencias.

Nota: esta tabla también admite otra interpretación, en la que en lugar de estudiar el sexo de los estudiantes en cada tipo de facultad, se estudie el tipo de estudios por sexo (es decir, podemos hacer la interpretación por filas).

Facultad / Sexo	Ciencias	Letras	Total
Hombre	13608 (52.94 %)	12096 (47.06 %)	25704 (100 %)
Mujer	16632 (67.35 %)	8064 (32.65 %)	24696 (100 %)
Total	30240 (60 %)	20160 (40 %)	50400 (100 %)

De este modo podríamos decir que en esta universidad, el 60 % de los alumnos estudian en facultades de Ciencias y el 40 % lo hace en facultades de Letras. Y por sexos, mientras los hombres se reparten en un 52.94 % en facultades de Ciencias y el 47.06 % restante en facultades de Letras, entre las mujeres las diferencias son mucho más acusadas ya que un 67.35 % estudian en facultades de Ciencias y solo un 32.65 % lo hace en facultades de Letras.

Tanto entre los hombres como entre las mujeres se mantiene la tendencia global y es mayor la proporción de los que estudian en las facultades de Ciencias que en las de Letras.

Otra forma, aunque menos común, de estandarizar por tamaño es la **razón**, que consiste en comparar, mediante un cociente, los casos que hay en una categoría con los que hay en otra categoría.

Si estamos interesados en conocer, en las facultades de Letras, la razón de hombres (12096) a mujeres (8064), construimos el cociente y simplificamos:

$$\text{razón} = \frac{12096}{8064} = \frac{3}{2}$$

es decir que, en las facultades de Letras, hay 3 hombres por cada 2 mujeres.

Para unificar la terminología, **las razones se suelen dar en unidades «por cada 100» unidades**. De este modo:

$$\text{razón} \times 100 = \frac{12096}{8064} \times 100 = \frac{3}{2} \times 100 = 150$$

es decir, que en las facultades de Letras, hay 150 hombres por cada 100 mujeres.

Si comparamos tipos de estudios de las mujeres de esta universidad, tendríamos:

$$\text{razón} \times 100 = \frac{16632}{8064} \times 100 = \frac{33}{16} \times 100 = \frac{825}{4} = 206.25$$

es decir, que en el grupo de las mujeres universitarias, hay 33 realizando estudios de Ciencias por cada 16 que realizan estudios de Letras. O bien, hay aproximadamente 206 mujeres en las facultades de Ciencias por cada 100 que están en las facultades de Letras.

Otro tipo de razones, que se usan más que las anteriores, son las **tasas**.

Todos hemos oído hablar de tasas de nacimiento, de mortalidad, de divorcios, etc... Así como en las razones se comparan el número de casos de un subgrupo o categoría con los de otro subgrupo, las tasas indican comparaciones entre el número de casos reales y el número de casos potenciales. Por ejemplo, para determinar la tasa de fecundidad en una determinada población se puede calcular el número de nacimientos vivos reales dividido por el número de mujeres en edad de quedarse embarazadas (que representan casos potenciales); o del mismo modo, la tasa de divorcios se calcula como el número de divorcios reales dividido por el número de matrimonios que ocurren en un período de tiempo (en un año, por ejemplo).

Las tasas suelen darse en términos de 1000 casos potenciales (es decir, se multiplica por mil el resultado del cociente). Por ejemplo, la tasa de natalidad en España en el año 2008 fue de 9.87 nacimientos por cada 1000 habitantes.

$$\text{tasa} = \frac{\text{frecuencia de casos reales}}{\text{frecuencia de casos potenciales}} \times 1000$$

Otro tipo de tasa muy utilizado es la **tasa de cambio o tasa de variación (porcentual)** que suele utilizarse para comparar un valor de una población en dos instantes diferentes de tiempo. Se suele expresar en porcentaje. Si un producto, en un año, ha pasado de costar 80 euros a costar 100 euros, la tasa de cambio sería:

$$\text{tasa de cambio} = \frac{\text{valor actual - valor origen}}{\text{valor origen}} \times 100 = \frac{100 - 80}{80} \times 100 = 25\,\%$$

el precio ha aumentado en un 25 % (respecto al valor original). Efectivamente, el 25 % de 80 euros son 20 euros, que es lo que ha aumentado el precio.

La tasa de cambio puede ser negativa, cuando el valor disminuye en el tiempo en lugar de aumentar.

Nota: es importante observar que si el precio original hubiera sido de 100 euros y pasa a costar 120 euros, la tasa de cambio sería:

$$\text{tasa de cambio} = \frac{\text{valor actual - valor origen}}{\text{valor origen}} \times 100 = \frac{120 - 100}{100} \times 100 = 20\,\%$$

en este caso la tasa de cambio sería del 20 %.

Es decir, que un mismo aumento de 20 euros, nos da distintas tasas de cambio, dependiendo del valor original.

2.4. Gráficos unidimensionales

Como ya hemos comentado, el primer paso en el análisis de los datos consiste en la reducción de la masa de datos para poder obtener una primera información acerca de las características del fenómeno que estamos estudiando.

Para hacernos una idea del comportamiento de una variable, además de las tablas de frecuencias, suele ser muy útil utilizar representaciones gráficas, que nos permiten visualizar, si las hay, características destacables.

Existen gráficos más o menos sofisticados, pero en general contienen la misma información. Vamos a comentar algunos de los más elementales.

Distinguiremos los gráficos dependiendo de si nuestra distribución está agrupada en intervalos o no.

2.4.1. Gráficos para distribuciones no agrupadas en intervalos

Retomemos el ejemplo que vimos al construir las tablas de frecuencias, en el que los valores observados de una variable, X, fueron los siguientes:

18	20	22	19	18
20	18	19	21	20
20	21	18	20	21
19	20	21	18	20

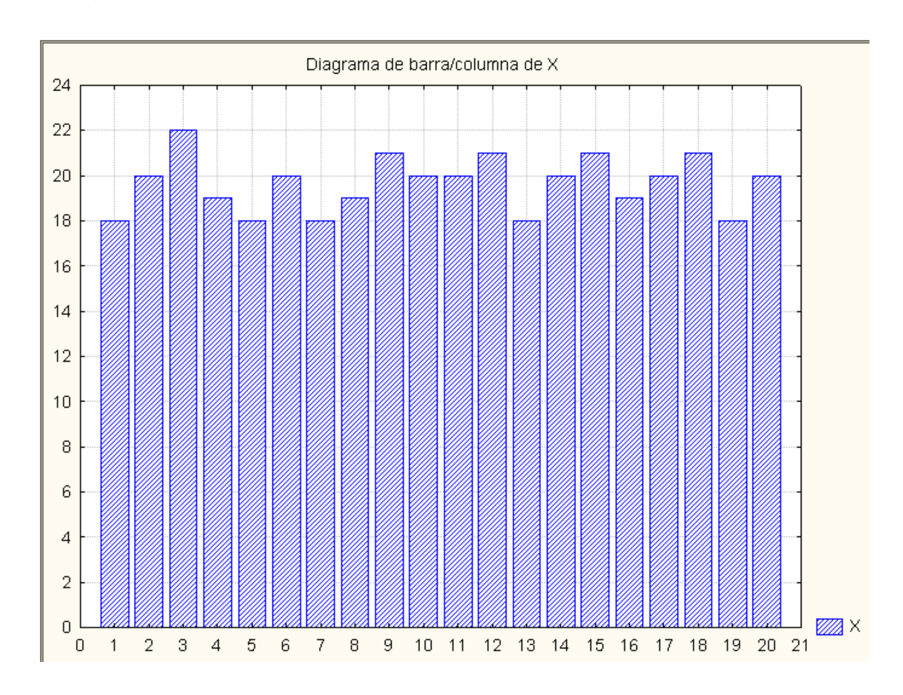
Para representar estos datos gráficamente, podemos utilizar:

Gráfico de barras:

Este tipo de gráfico, se utiliza para representar valores o frecuencias. Podemos representar:

- para cada caso, el valor observado de la variable.
- para cada valor de la variable, su frecuencia.

En la primera situación, para un sistema de ejes coordenados, dibujamos sobre el eje horizontal cada uno de los casos y levantamos, para cada uno de estos valores, una barra cuya altura será igual al valor observado de la variable en ese caso.



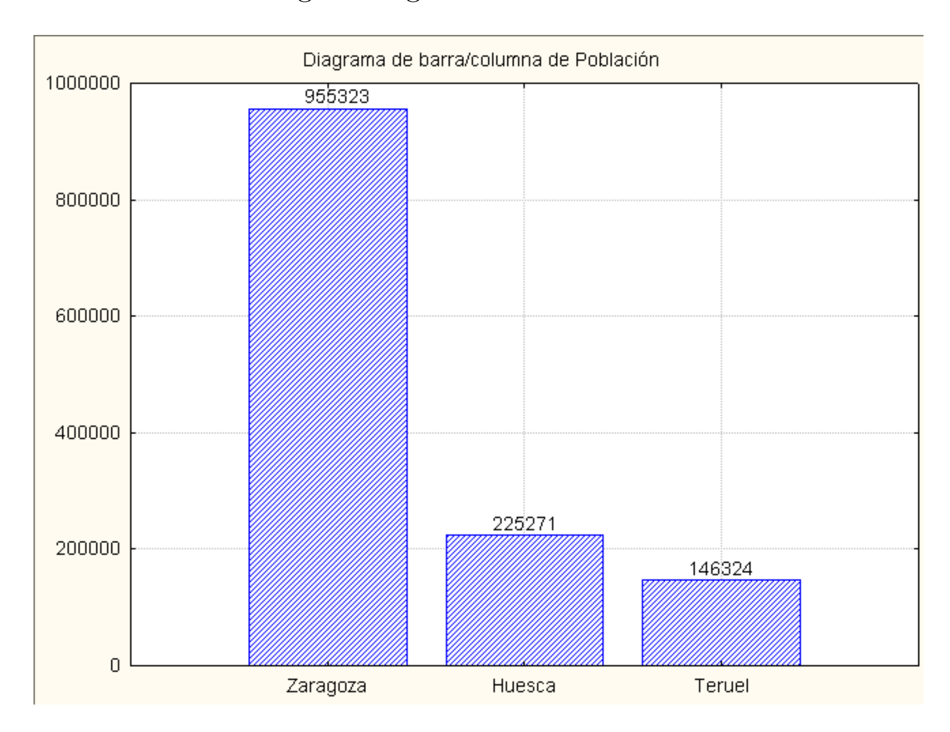
Se suele utilizar para representar valores de una variable cuando podemos identificar los casos.

Por ejemplo:

Si sabemos que la población de las provincias aragonesas (1 de enero de 2008) es la siguiente:

Provincia	Población
Zaragoza	955323
Huesca	225271
Teruel	146324

la podemos mostrar en el siguiente gráfico:



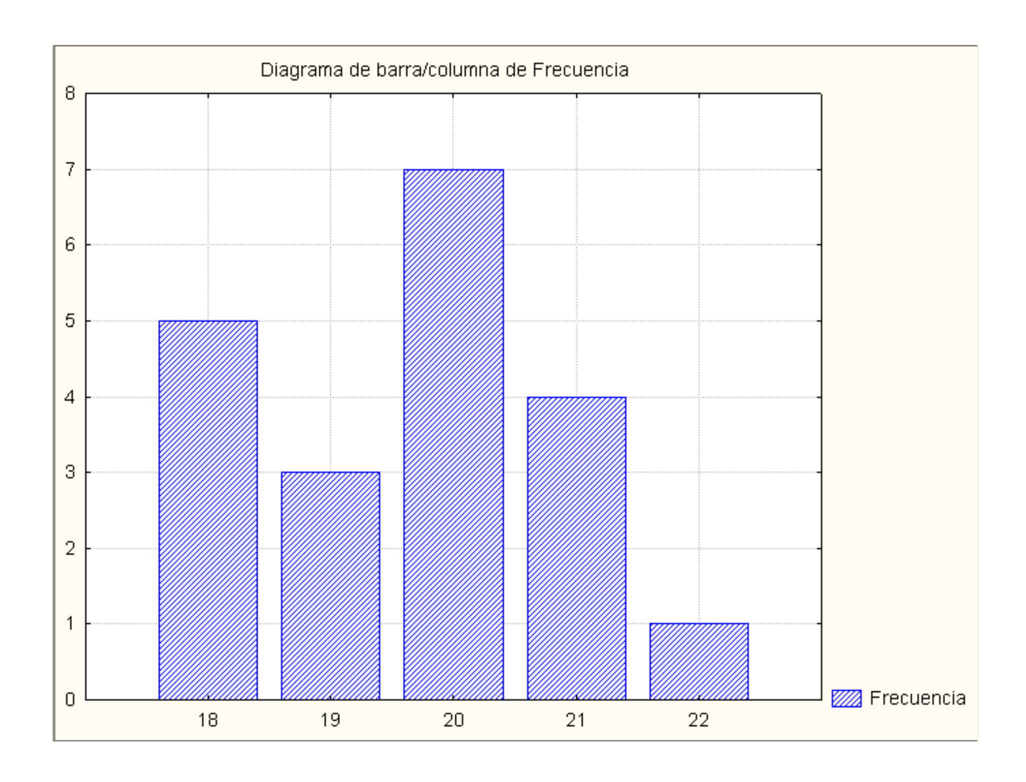
Este gráfico nos informa sobre los valores de la variable para cada caso, por lo que es interesante para **mostrar** la información, pero no sirve para resumir la información.

El **gráfico de barras** es uno de los más usados para **representar las frecuencias**:

Sobre un sistema de ejes dibujaremos en el eje horizontal los distintos valores de la variable y en el eje vertical la frecuencia de cada uno de ellos. Para cada valor de la variable, en el eje horizontal se levanta una barra cuya altura será igual a su frecuencia absoluta, o a la frecuencia absoluta acumulada.

Estos gráficos también se pueden hacer con los porcentajes –frecuencias relativas multiplicadas por 100–. En ese caso solo cambia la escala ya que la forma del gráfico queda exactamente igual.

x_i	n_i	N_i	f_i	F_i
18	5	5	5/20	5/20
19	3	8	3/20	8/20
20	7	15	7/20	15/20
21	4	19	4/20	19/20
22	1	20	1/20	1

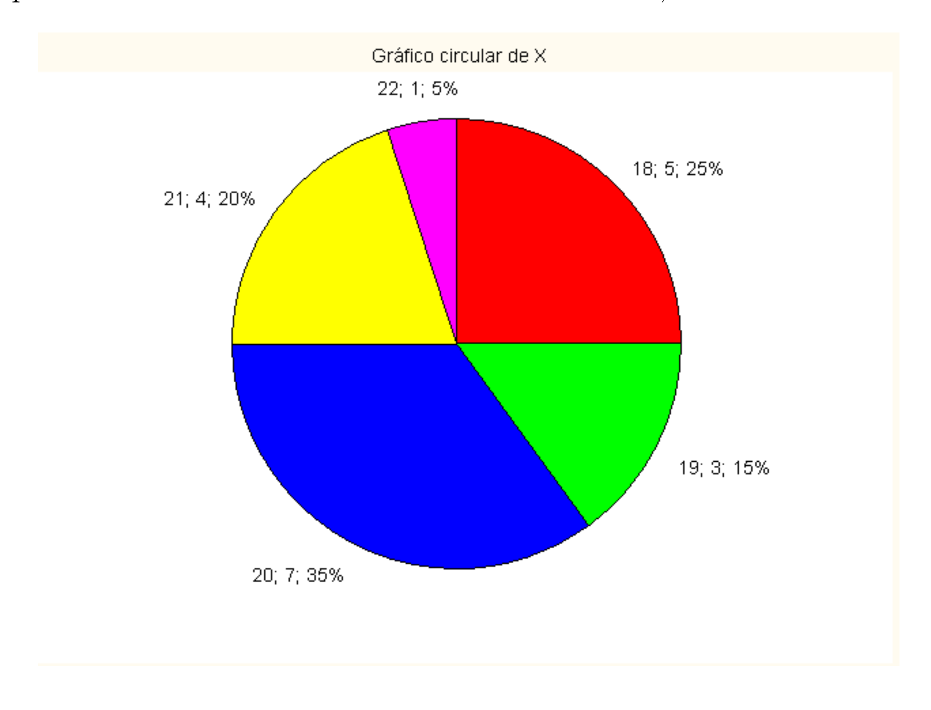


Para representar frecuencias relativas también es muy útil el:

Gráfico de sectores

Es un gráfico en el que el área de cada sector representa la frecuencia relativa de cada valor de la variable, respecto al total.

Es útil para visualizar las diferencias de las frecuencias, entre las distintas categorías.



2.4.2. Gráficos para distribuciones agrupadas

Histograma

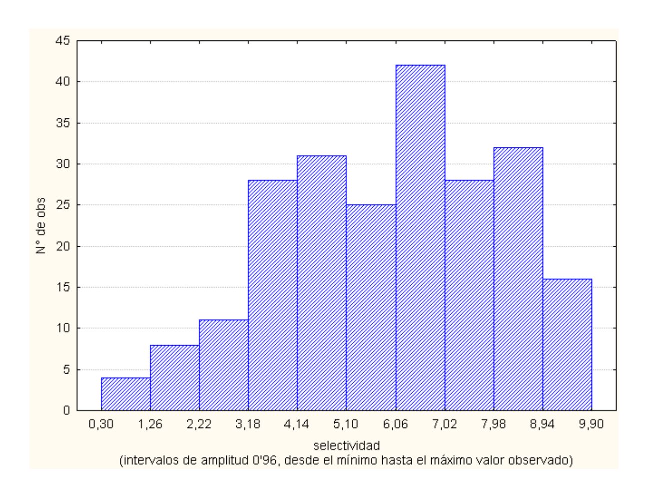
Cuando tenemos distribuciones agrupadas, sobre el eje horizontal se dibujan los intervalos y sobre cada uno de ellos se levanta un rectángulo cuya área sea proporcional a la frecuencia absoluta dentro del intervalo.

Como ya hemos comentado, el número de intervalos y su amplitud, quedan a criterio del investigador.

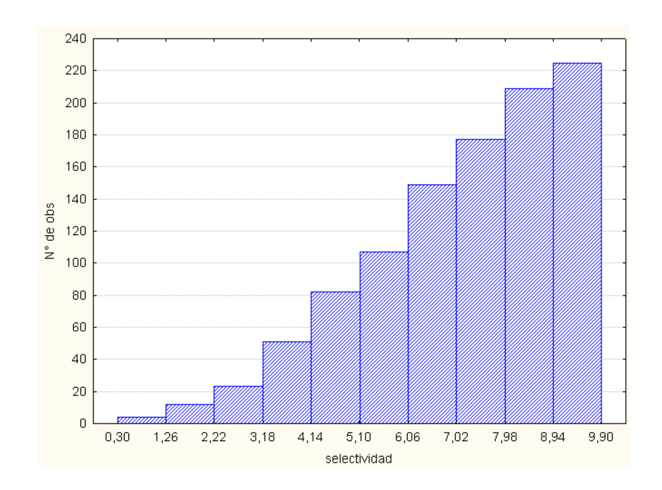
IMPORTANTE: si todos los intervalos tienen la misma amplitud, las alturas de los rectángulos pueden ser iguales a la frecuencia absoluta, pero si hay intervalos de distinta amplitud (a_i), entonces las alturas (h_i) se calculan dividiendo la frecuencia absoluta por la amplitud:

$$\mathrm{h}_i = \frac{n_i}{a_i}$$, que es lo que se llama **densidad de frecuencia**.

Por ejemplo, en el caso de las notas en Selectividad de 225 estudiantes (página 10):



También se pueden representar las frecuencias acumuladas con el Histograma de frecuencias acumuladas o gráfico de escalera:

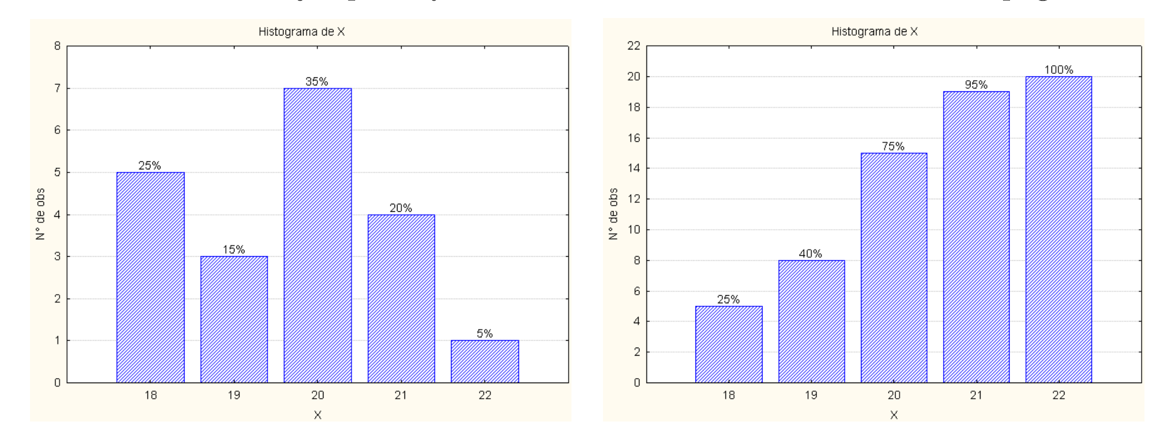


Este gráfico sólo tiene sentido para intervalos de la misma amplitud.

Histograma de datos categóricos

En algunas ocasiones el histograma también se utiliza para representar la frecuencia (absoluta, relativa o acumulada) de cada uno de los valores observados de la variable, como si fuese un gráfico de barras.

Para los datos del ejemplo cuya tabla de frecuencias construimos en la página 20:



2.5. Medidas de una variable cuantitativa

Como se ha comentado, para hacer manejable la masa de datos procedentes de la observación estadística, es necesario reducir el volumen de los datos; hemos visto que esto se puede conseguir construyendo la tabla de distribución de frecuencias.

En el caso de las variables **cuantitativas**, es posible reducir aún más estas distribuciones, valiéndonos de unos pocos números que *describan o caractericen* a las distribuciones de frecuencias. Estos números, que reciben el nombre de **características**, nos indican los rasgos más importantes de las distribuciones de frecuencias y se suelen clasificar en los siguientes grupos:

1. Medidas de posición. Estos a su vez se dividen en:
   - Centrales: media aritmética, mediana y moda.
   - No centrales: cuantiles.
2. Medidas de dispersión.
3. Medidas de asimetría.
4. Medidas de apuntamiento.

Vamos a analizarlas detenidamente.

2.6. Medidas de posición

Las medidas de posición, son unos valores alrededor de los cuales se agrupan los valores de la variable, y que nos resumen la posición de la distribución sobre el eje horizontal.

Existen dos tipos de medidas de posición: las centrales y las no centrales.

De las medidas de posición central o promedios, las más utilizadas son: la media aritmética, la mediana y la moda.

Las medidas de posición no central son los cuantiles.

2.6.1. La media aritmética

La media aritmética: se define como la suma de todos los valores observados de la distribución, dividida por el número total de observaciones.

Si agrupamos los valores que se repiten, la expresión de la media es:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{x_1 n_1 + \cdots + x_k n_k}{N}$$

Este es el promedio más utilizado en la práctica y esto es así por las ventajas que tiene y que son fundamentalmente:

- Tiene en cuenta todos los valores observados.
- Es fácil de calcular y tiene un claro significado estadístico.
- Es única.

Por otra parte tiene el inconveniente de la influencia que ejercen los valores extremos de la distribución sobre ella.

Propiedades

1. La suma de las desviaciones (diferencias con el correspondiente signo) de los valores de la variable, respecto a su media aritmética, es igual a cero.

   En efecto:
   $$\sum_{i=1}^{k}(x_i - \bar{x})n_i = \sum_{i=1}^{k} x_i n_i - \bar{x}\sum_{i=1}^{k} n_i = N\bar{x} - N\bar{x} = 0$$

2. Si tenemos que $u_i = a + bx_i$, siendo a y b valores cualesquiera, con b distinto de cero (lo que equivale a hacer un cambio de origen y escala), la media aritmética puede expresarse de la forma siguiente: $\bar{u} = a + b\bar{x}$

Comprobarlo es muy sencillo:

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{k} u_i n_i = \frac{1}{N}\sum_{i=1}^{k}(a + bx_i)n_i = a\frac{1}{N}\sum_{i=1}^{k} n_i + b\frac{1}{N}\sum_{i=1}^{k} x_i n_i =$$

$$= a\frac{N}{N} + b\frac{1}{N}\sum_{i=1}^{k} x_i n_i = a + b\bar{x}$$

Esta propiedad, eligiendo convenientemente los valores a y b, es de gran utilidad en muchos casos, para simplificar el cálculo de la media aritmética.

3. Si en una distribución de frecuencias se clasifican las observaciones en dos grupos mutuamente excluyentes, la media aritmética de todo el conjunto se relaciona con las medias aritméticas de los subconjuntos parciales, de la siguiente forma:

$$\bar{x} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2}{N}$$

donde:

- $\bar{x}$= media del conjunto total.
- N= número de observaciones del conjunto total.
- $\bar{x}_1$=media del primer subconjunto.
- N_1= número de observaciones del primer subconjunto.
- $\bar{x}_2$=media del segundo subconjunto.
- N_2= número de observaciones del segundo subconjunto.
- y naturalmente, se verifica que $N = N_1 + N_2$

Esta propiedad se puede generalizar para el caso de dividir la población total en p subconjuntos mutuamente excluyentes. Es decir:

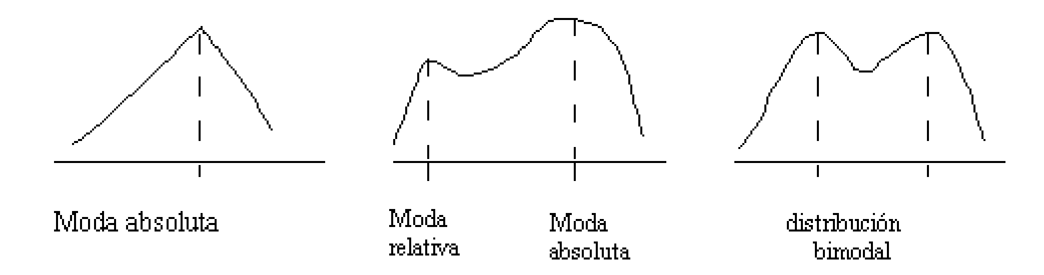
$$\bar{x} = \frac{\bar{x}_1 N_1 + \bar{x}_2 N_2 + \cdots + \bar{x}_p N_p}{N}$$

En donde se cumple que: $N = N_1 + N_2 + \cdots + N_p$

2.6.2. La moda

En una distribución, **la moda (Mo)** se define como «aquel valor de la variable cuya frecuencia no es superada por la frecuencia de ningún otro valor». Esta definición corresponde a la denominada moda absoluta. La moda relativa se define como «el valor de la variable cuya frecuencia no es superada por la de sus valores contiguos».

Puede darse el caso de que la máxima frecuencia corresponda a dos o más valores de la variable, en ese caso las distribuciones reciben el nombre de bimodales o multimodales.



En una distribución no agrupada en intervalos, la determinación de la moda absoluta y las modas relativas es inmediata.

En una distribución agrupada en intervalos la determinación con exactitud de la moda es imposible, por lo que se suelen hacer algunas suposiciones para poder calcular este valor.

Un criterio válido es tomar como moda la marca de clase del intervalo modal, entonces, para calcular la moda hay que hacer lo siguiente:

1º) Determinar cuál es el intervalo modal. El intervalo modal es aquel que tiene mayor densidad de frecuencia (en el caso de que todos los intervalos tengan la misma amplitud coincide con el intervalo en el que hay mayor frecuencia absoluta).

2º) La moda será, entonces, la marca de clase de este intervalo.

2.6.3. La mediana

Para una distribución discreta no agrupada en intervalos, se define **la mediana (Me)**, como el valor de la variable que ocupa el lugar central, supuestos ordenados los valores de menor a mayor. También se puede definir como el valor de la variable que divide a la distribución en dos partes con el mismo número de observaciones.

Si el número de observaciones es impar, entonces el valor de la mediana es inmediato (el valor que ocupe el lugar $\frac{N+1}{2}$).

Si el número de datos es par, suele tomarse como valor de la mediana, la media aritmética de los dos valores centrales, es decir, de los que ocupan los lugares $\frac{N}{2}$ y $\frac{N}{2}+1$. Naturalmente cuando estos dos valores son iguales, la mediana coincide con el valor común.

En el supuesto de una distribución agrupada en intervalos, la mediana será alguno de los valores contenidos en el intervalo al que corresponda una frecuencia acumulada **inmediatamente superior** a $\frac{N}{2}$; el cual se denomina **intervalo mediano**.

No podemos determinar exactamente cuál de los valores del intervalo es la mediana, y se pueden seguir varios criterios para elegir uno de ellos. Por simplificar nosotros tomaremos como mediana, la marca de clase del intervalo mediano.

Propiedad:

La mediana no depende de los valores extremos y por tanto, puede calcularse aún cuando estos se desconozcan; basta con conocer su frecuencia.

Ejemplos:

x_i	n_i	N_i
1	10	10
2	12	22
3	7	29
4	7	36
5	3	39

$\frac{39+1}{2} = 20$ luego, Me=2

x_i	n_i	N_i
1	10	10
2	12	22
3	7	29
4	8	37
5	3	40

$\frac{40}{2} = 20$ y $\frac{40}{2} + 1 = 21$

luego, Me=2

x_i	n_i	N_i
1	10	10
2	10	20
3	7	27
4	8	35
5	5	40

$\frac{40}{2} = 20$ y $\frac{40}{2} + 1 = 21$

luego, Me=$\frac{2+3}{2} = 2.5$

$(l_{i-1}, l_i]$	n_i	N_i
10-11	10	10
11-12	12	22
12-13	12	34
13-14	10	44
14-15	7	51

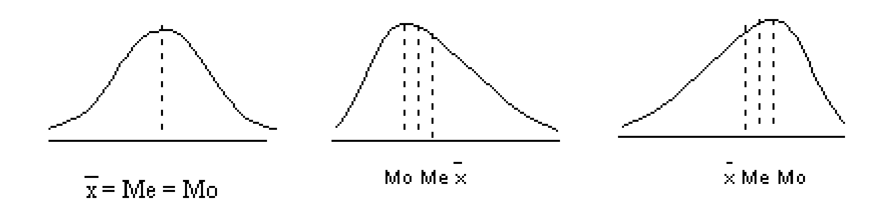
$\frac{51}{2} = 25.5 \Rightarrow$ Intervalo mediano 12-13

Me=marca de clase del intervalo mediano =$\frac{12+13}{2} = 12.5$

Relación entre la media aritmética, la mediana y la moda

En realidad estos tres promedios no deben emplearse de forma excluyente. Cada uno tiene su significado y se relacionan con aspectos diferentes de la distribución. No obstante existe cierta relación entre ellos que es conveniente saber.

En las distribuciones de frecuencias Normales (se estudiará más adelante), coinciden exactamente los tres promedios. **Si la distribución es acampanada pero no presenta simetría, la mediana está situada entre la moda y la media aritmética.**



Si la asimetría es a la derecha: Mo < Me < $\bar{x}$

Si la asimetría es a la izquierda: $\bar{x}$ < Me < Mo

2.6.4. Medidas de posición no central

Además de las medidas de posición centrales vistas hasta ahora, existen otros valores, no centrales, que pueden considerarse como indicadores de una determinada posición en la distribución.

Estos valores, llamados generalmente **cuantiles**, constituyen una generalización del concepto de la mediana.

Así como la mediana divide a la distribución en dos partes, cada una con el mismo número de observaciones que la otra, si dividimos la distribución en cuatro partes, cada una de ellas con el mismo número de observaciones, obtendremos tres valores, que se denominan **cuartiles**.

Análogamente, si dividimos la distribución en diez partes con el mismo número de observaciones, obtendremos nueve valores, que se denominan **deciles**. Y si la dividimos en cien partes, los correspondientes noventa y nueve valores se denominan **percentiles**.

En general, los $q-1$ valores que dividen a la distribución en q partes con el mismo número de observaciones se denominan **cuantiles de orden** q.

La determinación de los cuantiles en una distribución no agrupada en intervalos, es análoga a la de la mediana.

- En general, el r-ésimo cuantil de orden q, será aquel valor de la variable al cual corresponde una frecuencia acumulada inmediatamente superior a $\frac{rN}{q}$.

- Por ello, el cuantil r-ésimo de orden q será el valor de la variable que ocupa el lugar $\frac{r}{q}(N-1)+1$.

Cuando se trate de una distribución agrupada en intervalos, ya sabemos que no vamos a poder calcular el valor exacto, pero lo podemos aproximar tomando la marca de clase del intervalo en el que se encuentra el cuantil correspondiente, y que es el primer intervalo con frecuencia acumulada mayor que $\frac{rN}{q}$.

Por ejemplo, los valores que encierran el 70 % central de la distribución serán los percentiles: 15 y 85.

2.7. Medidas de dispersión

Las medidas de posición que acabamos de estudiar tienen como misión, no solo situar la distribución en el eje real, sino además sintetizar la información que proporciona la distribución.

El promedio con el que representamos una distribución llevará a cabo esta misión con mayor o menor fidelidad dependiendo de la relación que exista entre los valores de la variable y el promedio.

Así, si todos los valores fueran iguales, la media, por ejemplo, coincidiría con todos ellos por lo que representaría fielmente a la distribución.

A medida que los valores individuales de la variable difieran del promedio, la representatividad de este será cada vez menor.

Por ello, para evaluar la representatividad de un promedio, necesitamos un indicador que, de alguna forma, nos cuantifique el grado de separación de los valores de la variable respecto al promedio en cuestión.

En este apartado estudiaremos las medidas de dispersión. Hay que tener en cuenta que existen dos tipos de medidas de dispersión: las absolutas y las relativas.

2.7.1. Medidas de dispersión absoluta

Con las medidas de dispersión absoluta se trata de medir la separación que, por término medio, existe entre los distintos valores de la variable, por lo que serán medidas que vendrán expresadas en la misma clase de unidades que la variable.

Las principales medidas de dispersión absoluta son:

El recorrido o rango

El **recorrido o rango o amplitud** se define como la diferencia entre el mayor y el menor valor de la variable. Es decir : $\text{Re} = \text{máx}\, x_i - \text{mín}\, x_i = x_k - x_1$

Si tenemos dos conjuntos de individuos en los que estamos estudiando la característica «peso», si el recorrido del primer conjunto es Re(1)= 10 kg y el del segundo es Re(2)=5 kg, podemos considerar que la primera población tiene mayor dispersión absoluta que la segunda.

El inconveniente de esta medida es que solo tiene en cuenta los valores extremos.

La varianza

De todas las medidas de dispersión absoluta, la varianza y su raíz cuadrada, la desviación típica, son las más importantes.

Hasta ahora, al hablar de dispersión absoluta, no nos hemos referido a la solución que parece más simple: promediar las desviaciones respecto a la media aritmética, con el signo correspondiente. Es decir, considerar la suma $\frac{\sum_{i=1}^{k}(x_i-\bar{x})}{N}$, pero como ya vimos en las propiedades de la media, esta suma es nula $\left(\frac{\sum_{i=1}^{k}(x_i-\bar{x})}{N}=0\right)$ y es por esto por lo que no podemos utilizarla como medida de dispersión.

Ahora bien, si esta suma es igual a cero es porque las desviaciones positivas compensan exactamente las negativas, por lo que, podemos eliminar el problema utilizando una potencia par de las desviaciones.

De todas las potencias pares, elegimos la más sencilla, y surge así la nueva medida de dispersión denominada varianza, que definimos como la media aritmética de los cuadrados de las desviaciones de los valores observados de la variable respecto a la media aritmética de la distribución. Se representa por S'^2 y es:

$$\mathrm{S}'^2=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{N}$$

Evidentemente, el valor numérico de S'^2 describe el mayor o menor grado de dispersión de la distribución de frecuencias que se considere.

En general, cuanto más dispersas sean las observaciones, mayores serán las desviaciones respecto a su media, y mayor por tanto, el valor numérico de la varianza.

Propiedades:

1. La varianza nunca puede ser negativa: $\mathrm{S}'^2 \geq 0$

   Es evidente ya que es una media de cuadrados.

2. La varianza se puede calcular (desarrollando la expresión anterior) como:

$$\mathrm{S}'^2=\frac{\sum_{i=1}^{k}x_i^2 n_i}{N}-\bar{x}^2$$

3. La varianza no se altera ante un cambio de origen. Es decir, que si hacemos el cambio: $u_i = x_i + a$, la varianza de la variable U, es la misma que la de la variable X.

   En efecto: como $u_i = x_i + a$, sabemos que: $\bar{u} = \bar{x} + a$ y por lo tanto: $u_i - \bar{u} = x_i - \bar{x}$

   Elevando al cuadrado, multiplicando por n_i, sumando para todos los valores de i y dividiendo por N, tenemos que:

$$\mathrm{S}'^2(U)=\frac{\sum_{i=1}^{k}(u_i-\bar{u})^2 n_i}{N}=\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 n_i}{N}=\mathrm{S}'^2(X)$$

4. Si se hace un cambio de origen y de escala, es decir, si se efectúa el cambio de variable: $u_i = a + bx_i$, las varianzas de las dos variables están relacionadas por la expresión: $\mathrm{S}'^2(U) = b^2\mathrm{S}'^2(X)$.

En efecto: si hacemos un cambio de origen y de escala tenemos que: $u_i = a + bx_i$ y entonces $\bar{u} = a + b\bar{x}$, restando ambas igualdades tenemos que : $u_i - \bar{u} = b(x_i - \bar{x})$.

Elevando al cuadrado, multiplicando por n_i, sumando para todos los valores de i y dividiendo por N, se obtiene como resultado:

$$\mathrm{S}'^2(U) = \frac{\sum_{i=1}^{k}(u_i - \bar{u})^2 n_i}{N} = \frac{\sum_{i=1}^{k} b^2(x_i - \bar{x})^2 n_i}{N} = b^2 \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N} = b^2 \mathrm{S}'^2(X)$$

Ejemplos de cálculo de la varianza:

x_i	x_i^2
0	0
1	1
2	4
3	9
4	16
10	30

Cinco observaciones que no se repiten (frecuencias iguales a 1)

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{\sum_{i=1}^{k} x_i}{N} = \frac{10}{5} = 2$$

$$\mathrm{S}'^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2 = \frac{\sum_{i=1}^{k} x_i^2}{N} - \bar{x}^2 = \frac{30}{5} - 2^2 = 2$$

x_i	n_i	$x_i n_i$	$x_i^2 n_i$
0	2	0	0
1	3	3	3
2	1	2	4
3	4	12	36
	10	17	43

Diez observaciones, con cuatro valores distintos que se repiten

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{17}{10} = 1.7$$

$$\mathrm{S}'^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} - \bar{x}^2 = \frac{43}{10} - 1.7^2 = 1.41$$

La desviación típica

La varianza de la variable viene expresada en unidades de distinto orden que la variable a la que se refiere. Así, si la variable se refiere a la estatura, expresada en centímetros, la varianza será un cierto número expresado en centímetros cuadrados. Esta es la razón por la que, para obtener una medida de dispersión, pero expresada en las mismas unidades que la variable, se emplea la **desviación típica o desviación estándar**, que es igual a la **raíz cuadrada de la varianza, con signo positivo**. Se representa por S':

$$\mathrm{S}' = +\sqrt{\mathrm{S}'^2} = +\sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N}}$$

Al venir expresada en las mismas unidades que la variable, permite su comparación con los valores de la variable.

Las propiedades de la desviación típica se deducen fácilmente de las de la varianza.

La cuasivarianza

Es una medida muy similar a la varianza (la única diferencia para el cálculo está en el denominador):

$$\mathrm{S}^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N-1}$$

y es muy utilizada en Inferencia Estadística.

La cuasidesviación típica

Es la raíz cuadrada positiva de la cuasivarianza.

$$\mathrm{S} = +\sqrt{\mathrm{S}^2} = +\sqrt{\frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N-1}}$$

2.7.2. Medidas de dispersión relativa

Con las medidas de dispersión relativa, se trata de medir la dispersión, con independencia de la clase de unidades en que venga expresada la variable. Estas medidas, permiten comparar la dispersión existente en dos distribuciones, cuyas variables vengan expresadas en distinta clase de unidades.

De entre las medidas de dispersión relativa, llamadas también índices de dispersión, las más importantes son:

El recorrido relativo

Se define como el cociente entre el recorrido de la variable y la media aritmética:

$$\mathrm{Rr} = \frac{\mathrm{Re}}{|\bar{x}|}$$

Nos indica el número de veces que el recorrido contiene a la media aritmética.

El coeficiente de variación o índice de dispersión de Pearson

Es el más empleado de los índices de dispersión relativos. Se designa por CV:

$$\mathrm{CV} = \frac{\mathrm{S}'}{|\bar{x}|}$$

Este número nos indica el número de veces que la desviación típica contiene a la media, o lo que es lo mismo, el tanto que representa S' por cada unidad de $\bar{x}$ (es un tanto por

uno). También se puede interpretar como la expresión de S′ empleando como unidad de medida la media aritmética.

Tanto en la determinación de la desviación típica como en la de la media, se utilizan todos los valores de la distribución, por lo que es el índice más completo de los que venimos estudiando.

Puesto que el valor mínimo que puede tomar S′ es cero, este es también el mínimo valor (en valor absoluto) que puede tomar el coeficiente de variación y que corresponde al caso de máxima representatividad de la media aritmética.

Para dos distribuciones con igual dispersión absoluta, el coeficiente de variación es tanto menor cuanto mayor sea la media aritmética.

Ej.: Desviación típica de 2 kg en el peso de un bebé (media 7 kg) y en el peso de un adulto (media 75 kg). Los coeficientes de variación son, respectivamente, 2/7 y 2/75.

Por último, debemos hacer notar que este coeficiente no está definido cuando la media aritmética de la distribución es igual a cero.

2.8. Medidas de forma

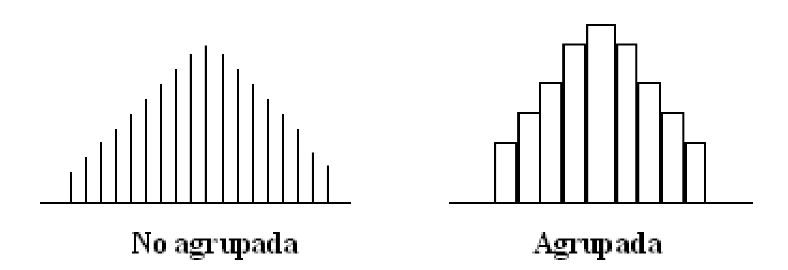
Ahora vamos a completar un poco más el análisis de una distribución, ya que con el estudio hecho hasta ahora, lo que hacemos es globalizar el comportamiento de una variable en un promedio y en la dispersión respecto a ese promedio, dejando de lado toda la disparidad, es decir, toda la variedad del comportamiento de la variable, fuera de la media.

Esta variedad se pone de manifiesto cuando representamos gráficamente la distribución.

Pues bien, en este apartado nos vamos a referir a ciertas medidas que nos van a dar una idea de la **forma** de la distribución, sin necesidad de realizar su representación gráfica.

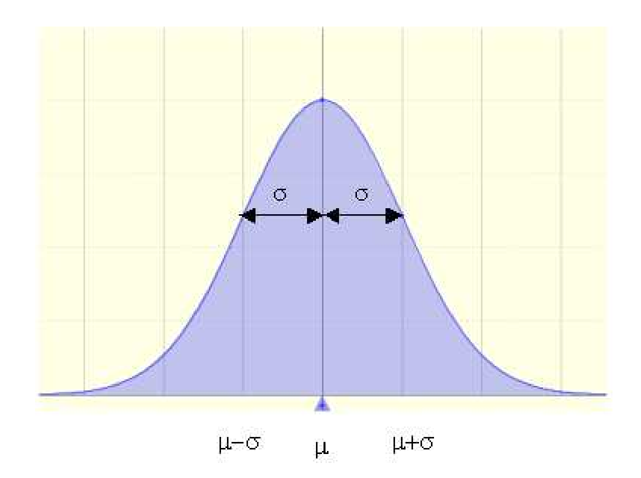
La forma de una distribución de frecuencias puede ser muy variada.

En una distribución campaniforme simétrica coinciden la media, la mediana y la moda y estas condiciones sugieren distribuciones cuyas frecuencias absolutas o relativas den lugar a representaciones del tipo:



Una curva continua, que puede servir como modelo matemático de ambos casos, es la

curva Normal de Gauss, que tiene la forma siguiente:



y cuyas características son:

1. Es simétrica.
2. Me = Mo = $\bar{x}$
3. Tiene como expresión:
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
donde μ es la media y σ es la desviación típica.

Evidentemente, no existe en la realidad ninguna variable cuya distribución de frecuencias relativas dé lugar a una curva así, pero se puede construir un modelo matemático o distribución de probabilidades con las propiedades citadas.

Dicho modelo es la distribución NORMAL cuya representación gráfica es la curva de Gauss y éste será el modelo de comparación para la simetría y la curtosis de cualquier distribución de frecuencias.

2.8.1. Medidas de simetría y asimetría

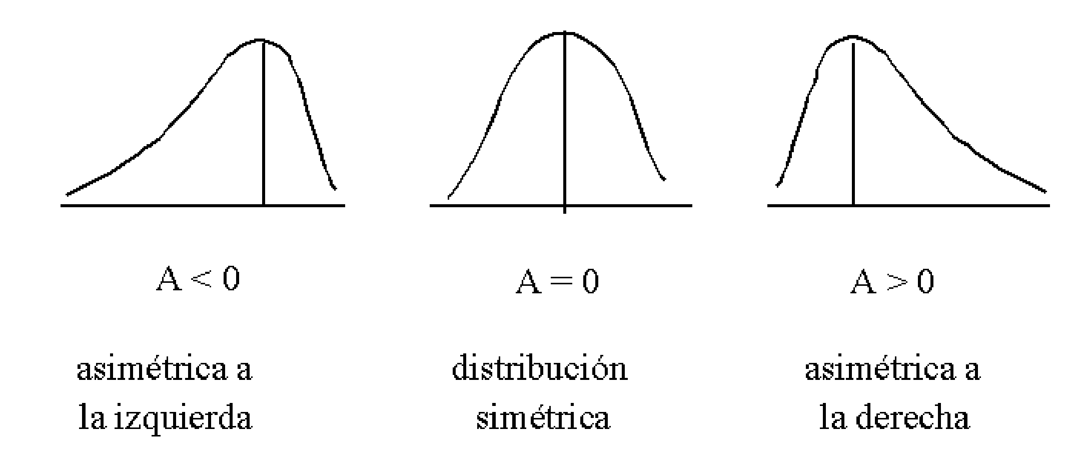
Las medidas de simetría nos permiten establecer un indicador del grado de simetría o asimetría que presenta la distribución, sin necesidad de llevar a cabo su representación gráfica.

Diremos que una distribución es simétrica cuando lo es su representación gráfica en coordenadas cartesianas. Es decir, que al trazar una recta paralela al eje de ordenadas por el punto x, existen el mismo número de valores x_i a ambos lados de dicha recta, equidistantes y a los que corresponde igual frecuencia.

Si la distribución es simétrica, el eje de simetría de su representación gráfica será una recta paralela al eje de ordenadas que pasa por el punto cuya abscisa es la media aritmética

(como puede comprobarse al recordar la primera propiedad de la media). Por ello, cuando la distribución es asimétrica, se suelen comparar los valores de la distribución con este promedio.

Existen varios coeficientes, A, que nos permiten determinar la simetría o el grado de asimetría de una distribución, pero para cualquiera de ellos la interpretación es la misma:



Un coeficiente de asimetría muy sencillo, aunque en algunos casos bastante impreciso, es el coeficiente de asimetría de Pearson:

Basándose en el hecho de que en una distribución simétrica unimodal se verifica que: $\bar{x} = \text{Mo} = \text{Me}$, Karl Pearson propuso como coeficiente de asimetría el siguiente:

$$\text{A}_P = \frac{\bar{x} - \text{Mo}}{\text{S'}}$$

Si la distribución presenta asimetría positiva, la media está desplazada a la derecha de la moda, por lo que se verifica que: $\bar{x} - \text{Mo} > 0$.

Por el contrario, si la distribución es asimétrica negativa, $\bar{x} - \text{Mo} < 0$.

Por lo tanto, el signo de A_P nos indica el de la asimetría:

- Si $\text{A}_P = 0$, la distribución es simétrica

- Si $\text{A}_P > 0$, la distribución es asimétrica positiva (a la derecha)

- Si $\text{A}_P < 0$, la distribución es asimétrica negativa (a la izquierda)

2.8.2. Medidas de curtosis o apuntamiento

Las medidas de curtosis se aplican a distribuciones campaniformes, es decir, unimodales simétricas o con ligera asimetría.

En esencia, las medidas de curtosis tratan de estudiar la distribución de frecuencias en la «zona central» de la distribución.

La mayor o menor **agrupación de frecuencias** alrededor de la media y en la zona central de la distribución, dará lugar a una distribución más o menos **apuntada**. Por esta razón a las medidas de curtosis se les llama también de «apuntamiento».

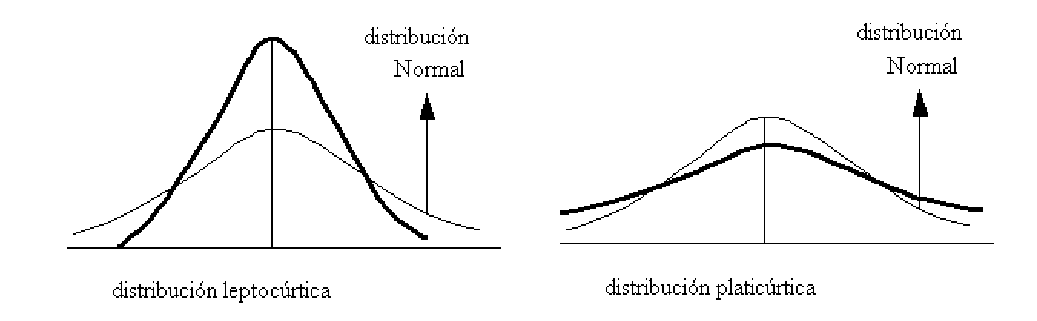
Con el coeficiente de curtosis se trata de medir el grado de apuntamiento de una distribución comparándolo con el de la distribución NORMAL.

Existen varios coeficientes que nos permiten calcular la curtosis, pero en todos los casos la interpretación es la misma:

- Si K = 0 , se dice que la distribución es mesocúrtica (normal).

- Si K > 0 , la distribución es leptocúrtica (más apuntada que la normal).

- Si K < 0 , la distribución es platicúrtica (menos apuntada que la normal).



2.9. Medidas de concentración

En el cálculo de la media aritmética, el numerador es la suma de todos los valores observados de la variable.

En muchos casos, dicho numerador no tiene sentido estadístico claro, por ejemplo, en una distribución de alturas, sería la suma de todas las alturas. Pero en otros casos, en particular cuando se trata de variables de carácter socio-económico, sí que lo tiene: así en una distribución de salarios, el numerador de la media aritmética representaría la masa total de salarios.

Pues bien, las medidas de concentración tienen por finalidad, precisamente, medir la uniformidad del reparto de dicha masa total.

Si todos los trabajadores perciben el mismo salario, la uniformidad de dicho reparto sería absoluta. Si, por el contrario, la masa total fuese percibida por un solo trabajador, entonces la falta de uniformidad sería total.

En general, las medidas de concentración tratan de poner de relieve el mayor o menor grado de igualdad en el reparto de la suma total de los valores de la variable.

Son, por tanto, **indicadores del grado de equidistribución** de la variable.

Para calcularlas habrá que relacionar el porcentaje acumulado de frecuencias (P_i) y el

porcentaje acumulado del total de la variable considerada (Q_i).

Supongamos la siguiente distribución de salarios mensuales en euros.

x_i	n_i	$x_i n_i$	$p_i = f_i$	$q_i = \frac{x_i n_i}{\sum x_i n_i}$	P_i	Q_i
800	10	8000	10/50=0.2	8000/80000=0.1	0.2	0.1
1500	20	30000	20/50=0.4	30000/80000=0.375	0.6	0.475
2000	15	30000	15/50=0.3	30000/80000=0.375	0.9	0.85
2400	5	12000	5/50=0.1	12000/80000=0.15	1	1
	50	80000				

(80000 es la masa salarial total).

Los datos de esta tabla representan:

- p_i, son las frecuencias relativas de trabajadores (proporción de trabajadores que están en cada grupo, respecto al total).

- P_i, son las frecuencias relativas acumuladas de trabajadores (proporción, respecto al total, de trabajadores acumulados comenzando por los que menos ganan).

- q_i, masa de salario que se reparte entre los miembros de la clase i-ésima, relativa a la masa salarial total.

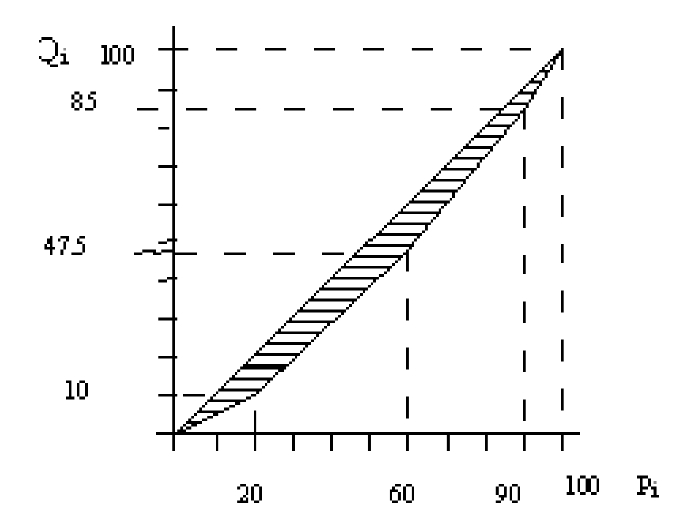
- Q_i, masa de salario acumulado hasta la clase i-ésima, comenzando por los que menos ganan, relativa a la masa salarial total.

Si ponemos en relación las columnas P_i y Q_i, obtenemos una información que nos indica el reparto de los salarios poniéndonos de relieve la concentración de los mismos.

En efecto, ordenados los trabajadores de menor a mayor salario, resulta que el 20 % de los trabajadores se reparte el 10 % de la masa salarial, el 60 % de los trabajadores el 47.5 % del dinero,...

2.9.1. La curva de Lorenz

La curva de Lorenz se obtiene representando los valores P_i y Q_i (P_i en el eje de abscisas y Q_i en el eje de ordenadas), y uniendo mediante líneas rectas cada punto (P_i, Q_i) con su consecutivo y además (P_1, Q_1) con el origen . La dibujamos en tantos por ciento.



Si el reparto hubiese sido equitativo, el 10 % de los trabajadores obtendría el 10 % de la masa salarial; el 20 % de los trabajadores, el 20 % de la masa salarial; etc... Es decir $P_i = Q_i, \forall i$, con lo cual la curva de Lorenz coincidiría con la diagonal.

Si, por el contrario, todos los trabajadores percibiesen el 0 % de la masa salarial excepto uno que percibiese el 100 %, entonces la curva de Lorenz estaría formada por los lados inferior y derecho.

En general, la curva de Lorenz es siempre convexa, y cuanto más convexa es, menos equitativa es la distribución (mayor concentración); mientras que cuanto más se aproxima a la diagonal, más equitativa es la distribución (menor concentración).

Si tomamos el área encerrada entre la diagonal y la curva, tenemos:

1) Si $P_i = Q_i, \forall i$, estamos en el caso de mínima concentración o máxima igualdad. Área = 0.

2) Si un solo trabajador percibiese el 100 % de la masa salarial, estaríamos en el caso de máxima concentración o mínima igualdad. Área $= \frac{1 \times 1}{2} = 0.5$

Por lo tanto, cuanto más se acerque a cero el área, tanto menor será la concentración y el grado de desigualdad existente en el reparto del total de la variable considerada.

2.9.2. Índice de concentración de Gini

Podemos construir un indicador del grado de concentración, comparando el área encerrada entre la curva de Lorenz y la diagonal, con el área del triángulo inferior. Esto es lo que hace el **Índice de Gini**.

Como el triángulo inferior tiene área 0.5, el índice de Gini es el doble del área comprendida entre la diagonal y la curva de concentración. ($\mathrm{I}_G = \frac{\text{Área}}{0.5} = 2 \times \text{Área}$).

Para evaluar el área indicadora del grado de concentración, basta con calcular el área de los distintos triángulos y rectángulos que se forman utilizando cualquier método conocido.

Se puede comprobar que el índice de Gini se obtiene exactamente mediante la expresión:

$$\mathrm{I}_G = 2\sum_{i=1}^{k} q_i \hat{P}_i - 1, \text{ donde } \hat{P}_i = P_i - \frac{p_i}{2}$$

Como el área está comprendida entre 0 y 0.5 esto significa que el índice de Gini está comprendido entre 0 y 1, lo que nos permite realizar la siguiente interpretación:

- $\mathrm{I}_G = 0$: No existe concentración (máxima equidad)
- I_G bajo (próximo a 0): Baja concentración (equidad elevada)
- I_G alto (próximo a 1): Alta concentración (poca equidad)
- $\mathrm{I}_G = 1$: Máxima concentración (nula equidad).

Veamos dos ejemplos de cómo calcularlo:

Ejemplo 1:

Con los datos del ejemplo anterior:

P_i	Q_i	p_i	q_i	$\hat{P}_i = P_i - \frac{p_i}{2}$	$q_i\hat{P}_i$
10/50	8000/80000	10/50	8000/80000	5/50	40000/(50 × 80000)
30/50	38000/80000	20/50	30000/80000	20/50	600000/(50 × 80000)
45/50	68000/80000	15/50	30000/80000	37.5/50	1125000/(50 × 80000)
50/50	80000/80000	5/50	12000/80000	47.5/50	570000/(50 × 80000)
					23335000/(50 × 80000)=0.58375

Entonces:

$$\mathrm{I}_G = 2\sum_{i=1}^{k} q_i \hat{P}_i - 1 = 2 \times \frac{23335000}{50 \times 80000} - 1 = 2 \times 0.58375 - 1 = 0.1675$$

Lo que significa que hay **poca concentración**. El área encerrada entre la curva de Lorenz y la diagonal, representa un 16.75 % del área del triángulo inferior. El reparto es **equitativo**.

Ejemplo 2:

Vamos a determinar si existe concentración en el reparto de los salarios que se dan en la siguiente tabla:

s_i	n_i	$s_i n_i$	p_i	q_i	P_i	$\hat{P}_i$	$q_i\hat{P}_i$
600	35	21000	35/90	21/102	35/90	17.5/90	367.5/(90 × 102)
1200	40	48000	40/90	48/102	75/90	55/90	2640/(90 × 102)
1800	10	18000	10/90	18/102	85/90	80/90	1440/(90 × 102)
3000	5	15000	5/90	15/102	90/90	87.5/90	1312.5/(90 × 102)
Sumas	90	102000					5760/(90 × 102)=0.629676

Entonces:

$$I_G = 2 \times 0.629676 - 1 = 0.259352$$

El valor obtenido indica que **existe concentración, aunque no muy grande**, ya que el área encerrada por la curva de Lorenz, representa un 25.94 % del área del triángulo inferior. El reparto de los salarios es bastante equitativo.

2.10. Ejemplo resuelto

Vamos a realizar un análisis estadístico de una variable unidimensional.

Para dirigir el análisis, a partir de los datos originales iremos respondiendo a una serie de preguntas utilizando los estadísticos adecuados que además deberemos interpretar correctamente.

Supongamos que tenemos la siguiente información respecto a los salarios (en euros) de los 1500 trabajadores de una gran empresa:

Salarios	Nº de trabajadores
$[1100, 1500]$	108
$(1500, 1700]$	377
$(1700, 1900]$	575
$(1900, 2100]$	351
$(2100, 2500]$	89

1. ¿Cuántos trabajadores cobran entre 1500 y 1700 euros?
2. ¿Qué porcentaje de los trabajadores de la empresa cobran entre 1900 y 2100 euros?
3. ¿Cuántos trabajadores cobran más de 1700 euros?
4. ¿Qué proporción representan los trabajadores que cobran hasta 1900 euros?
5. Dibuja un histograma que represente la distribución de los salarios de los trabajadores de esta empresa.
6. ¿Cuál es el salario más habitual en esta empresa?
7. ¿Qué salario no es superado por el 32.33 % de los trabajadores?
8. ¿Cuál es el salario medio de los trabajadores de esta empresa?
9. ¿Qué desviación típica tienen estos salarios?
10. La distribución de los salarios ¿es homogénea?
11. Si queremos utilizar el salario medio como representante de los salarios en esta empresa, ¿este salario medio es representativo?
12. Si nos dicen que, para los datos de esta empresa, el coeficiente de asimetría es 0.023 y que el coeficiente de curtosis es -0.120, ¿qué podemos decir respecto a la forma de la distribución?
13. El reparto de los salarios en esta empresa ¿es equitativo?
14. ¿Qué porcentaje de la masa salarial se reparten el 32.33 % de los trabajadores que menos ganan?

15. ¿Qué porcentaje de los trabajadores que menos ganan se reparten el 66.23 % de la masa salarial?

Solución.

Incluimos aquí una tabla cuyas columnas iremos construyendo a medida que las necesitamos para responder a las distintas preguntas.

Salarios	s_i	n_i	N_i	f_i	F_i	a_i	$h_i = \frac{n_i}{a_i}$	$s_i n_i$	$s_i^2 n_i$
$[1100, 1500]$	1300	108	108	0.072	0.072	400	0.27	140400	182520000
$(1500, 1700]$	1600	377	485	$0.251\hat{3}$	$0.32\hat{3}$	200	1.885	603200	965120000
$(1700, 1900]$	1800	575	1060	$0.383\hat{3}$	$0.70\hat{6}$	200	2.875	1035000	1863000000
$(1900, 2100]$	2000	351	1411	0.234	$0.940\hat{6}$	200	1.755	702000	1404000000
$(2100, 2500]$	2300	89	1500	$0.059\hat{3}$	1	400	0.2225	204700	470810000
		1500						2685300	4885450000

1. ¿Cuántos trabajadores cobran entre 1500 y 1700 euros?

   Nos preguntan cuántos trabajadores tienen un salario cuyo valor está dentro de este intervalo, por lo tanto son: 377 trabajadores.

2. ¿Qué porcentaje de los trabajadores de la empresa cobran entre 1900 y 2100 euros?

   Un porcentaje es una frecuencia relativa (proporción) multiplicada por 100, por lo tanto, serán

$$\frac{351}{1500} \times 100 = 0.234 \times 100 = 23.4\,\%$$

3. ¿Cuántos trabajadores cobran más de 1700 euros?

   Para responder a esta pregunta podemos utilizar las frecuencias absolutas acumuladas. Los trabajadores que cobran más de 1700 euros son todos menos los que cobran un salario menor o igual a dicha cantidad.

$$N - N_2 = 1500 - 485 = 1015$$

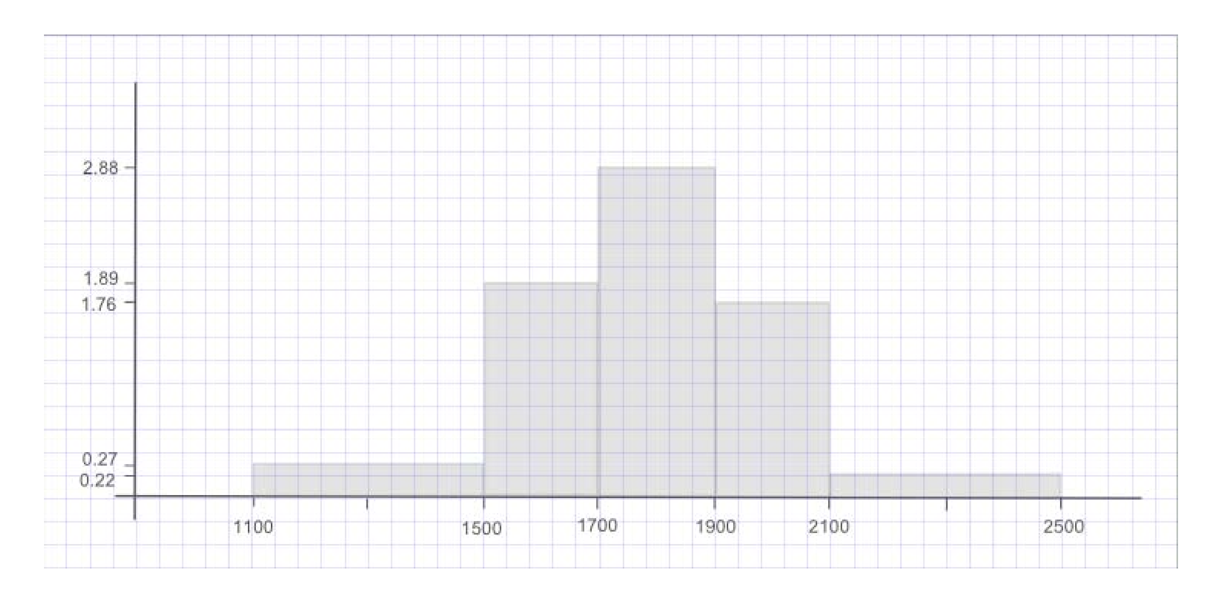
4. ¿Qué proporción representan los trabajadores que cobran hasta 1900 euros?

   Nos piden la proporción (frecuencia relativa) acumulada de trabajadores cuyos salarios no superan los 1900 euros.

$$F_3 = 0.70\hat{6}$$

5. Dibuja un histograma que represente la distribución de los salarios de los trabajadores de esta empresa.

   En un histograma las áreas de los rectángulos deben ser proporcionales a las frecuencias absolutas correspondientes. En nuestro caso los intervalos tienen distinta amplitud, por lo que las alturas deben ser las densidades de frecuencia (de este modo: área= $a_i h_i = n_i$)



6. ¿Cuál es el salario más habitual en esta empresa?

   Nos están preguntando por la moda. Primero buscamos el intervalo modal, que es aquel en el que hay mayor densidad de frecuencia: $(1700, 1900]$. Entonces, según el criterio que estamos utilizando, la moda será la marca de clase de este intervalo: Mo=1800.

7. ¿Qué salario no es superado por el 32.33 % de los trabajadores?

   Buscamos el salario tal que el porcentaje de trabajadores con un salario inferior es del 32.33 %.

   Tenemos que la frecuencia relativa acumulada hasta 1700 es 0.3233, por lo tanto, el salario que estamos buscando es: 1700 euros.

8. ¿Cuál es el salario medio de los trabajadores de esta empresa?

   Para calcular la media, como tenemos intervalos y necesitamos utilizar valores concretos de la variable, utilizaremos las marcas de clase (son valores que representan a todas las observaciones que se encuentran en cada intervalo). Entonces:

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{k} s_i n_i = \frac{1}{1500} 2685300 = 1790.2 \text{ euros}$$

9. ¿Qué desviación típica tienen estos salarios?

   Procediendo de forma análoga al cálculo de la media, calculamos primero la varianza:

$$\mathrm{S}'^2 = \frac{1}{N}\sum_{i=1}^{k} s_i^2 n_i - \bar{s}^2 = \frac{1}{1500} 4885450000 - (1790.2)^2 = 52150.62\hat{6}$$

   entonces, la desviación típica es:

$$\mathrm{S}' = \sqrt{52150.62\hat{6}} = 228.365117 \text{ euros}$$

10. La distribución de los salarios ¿es homogénea?

    Para estudiar la homogeneidad (lo parecidos que son entre sí los salarios), estudiaremos la dispersión relativa.

$$\text{CV} = \frac{S'}{\bar{s}} = \frac{228.3651}{1790.2} = 0.12756$$

    El coeficiente de variación es muy pequeño (está muy próximo a cero), por lo que existe muy poca dispersión relativa. Eso indica que la distribución de los salarios es muy homogénea.

11. Si queremos utilizar el salario medio como representante de los salarios en esta empresa, ¿este salario medio es representativo?

    Sí, porque al haber poca dispersión relativa esto significa que los salarios tienen poca dispersión respecto a la media. Es decir que son muy parecidos entre sí y parecidos a la media. Por lo tanto, podemos usar la media como representante de los salarios de la empresa.

12. Si nos dicen que, para los datos de esta empresa, el coeficiente de asimetría es 0.023 y que el coeficiente de curtosis es -0.120, ¿qué podemos decir respecto a la forma de la distribución?

    Hemos visto, al hacer el histograma que la distribución es campaniforme, por lo tanto, con estos coeficientes podemos añadir que también es ligeramente asimétrica a la derecha (muy poco) y ligeramente platicúrtica.

13. El reparto de los salarios en esta empresa ¿es equitativo?

    Para analizar si un reparto es equitativo tenemos que estudiar si existe concentración en el reparto, y para ello tendremos que calcular el índice de Gini o dibujar la curva de Lorenz. Vamos a hacer ambas cosas.

$$\text{I}_G = 2\sum_{i=1}^{k} q_i\hat{P}_i - 1$$

    Vamos a construir una tabla con los cálculos:

s_i	n_i	$s_i n_i$	p_i	q_i	P_i	$\hat{P}_i$	$q_i\hat{P}_i$
1300	108	140400	108/N	140400/S	108/N	54/N	$7581600/(N \times S)$
1600	377	603200	377/N	603200/S	485/N	296.5/N	$178848800/(N \times S)$
1800	575	1035000	575/N	1035000/S	1060/N	772.5/N	$799537500/(N \times S)$
2000	351	702000	351/N	702000/S	1411/N	1235.5/N	$867321000/(N \times S)$
2300	89	204700	89/N	204700/S	1500/N	1455.5/N	$297940850/(N \times S)$
Sumas	1500=N	2685300=S					$2151229750/(N \times S)$

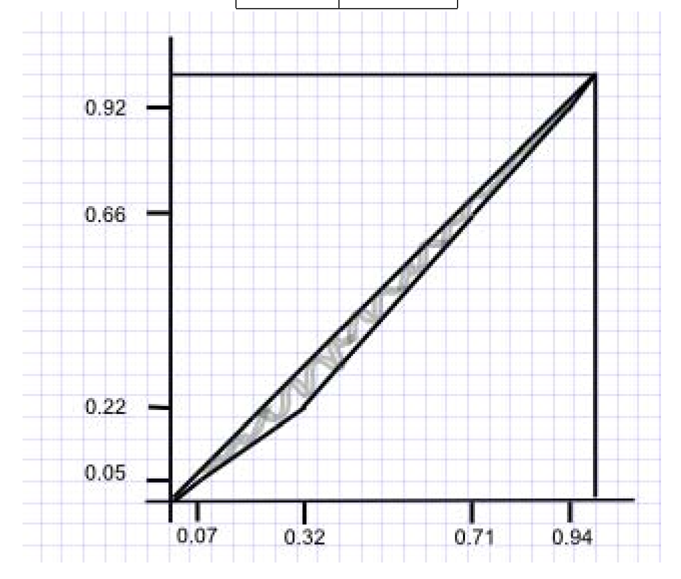
    Entonces:

$$\text{I}_G = 2 \times \frac{2151229750}{1500 \times 2685300} - 1 = 0.06815$$

    Hay muy poca concentración. El reparto es muy equitativo.

Vamos a resolverlo dibujando la curva de Lorenz. Construimos una tabla con los valores de P_i y de Q_i y dibujamos el gráfico correspondiente.

P_i	Q_i
0.072	0.0523
$0.32\hat{3}$	0.2246
$0.70\hat{6}$	0.6623
0.941	0.9238
1	1



Como vemos, el área que queda entre la curva y la diagonal es muy pequeña (el índice de Gini nos indica que es menos de un 7 % del área del triángulo inferior), por lo tanto hay muy poca concentración. El reparto es muy equitativo.

14. ¿Qué porcentaje de la masa salarial se reparten el 32.33 % de los trabajadores que menos ganan?

    Esta pregunta se puede responder con la tabla que hemos construido en el apartado anterior ya que nos piden relacionar proporciones acumuladas de trabajadores (P_i) y proporciones acumuladas de masa salarial (Q_i)

    El 32.33 % de los trabajadores que menos ganan corresponden a $P_i = 0.32\hat{3}$, y a ellos les corresponde una masa salarial acumulada de $Q_i = 0.2246$, es decir, un 22.46 % de la masa salarial total.

15. ¿Qué porcentaje de los trabajadores que menos ganan se reparten el 66.23 % de la masa salarial?

    Esta pregunta se responde de forma análoga a la anterior.

    El 66.23 % de la masa salarial ($Q_i = 0.6623$) corresponde a una proporción acumulada de trabajadores $P_i = 0.70\hat{6}$, es decir, al $70.\hat{6}$ % de los trabajadores que menos ganan.

Tema 3

Números índices

En el tema anterior vimos cómo describir perfectamente una variable cuantitativa, o lo que es lo mismo, cómo describir para la población estudiada, su comportamiento respecto a una determinada característica. Sin embargo, hay situaciones que por sus características particulares no se pueden describir de este modo. Esto es lo que ocurre, por ejemplo, con la evolución de las magnitudes económicas.

Ciertas magnitudes económicas suelen variar en el tiempo o en el espacio (niveles de precios, de salarios, etc...), por lo que surge la necesidad de cuantificar estas variaciones para así disponer de una medida objetiva de la importancia de dichas variaciones.

Así, por ejemplo, nos puede interesar la variación en los precios de la vivienda y analizar cómo ha ido variando a lo largo de los últimos años, respecto a un año concreto. O nos interesa la evolución de los salarios a partir, por ejemplo, de 2008.

Por otra parte, en unas ocasiones nos interesarán las variaciones de una magnitud (precio de la gasolina) y en otras las de un conjunto de magnitudes (precio de los carburantes). Pues bien, esta es la cuestión que vamos a tratar en este tema.

Definición: Llamaremos **números índices** a unas medidas estadísticas que sirven para comparar una magnitud o un grupo de magnitudes en dos situaciones, una de las cuales se considera de referencia. La comparación se puede efectuar en el tiempo o en el espacio.

Los índices que vamos a estudiar van a referirse generalmente a la evolución de una magnitud en el tiempo. Así, a la situación inicial la llamaremos PERÍODO BASE O PERÍODO DE REFERENCIA y a la situación que queremos comparar PERÍODO ACTUAL.

La teoría de los números índices se ha desarrollado fundamentalmente para el estudio de las variaciones de los precios, precisamente para tratar de medir el nivel general de precios, e inversamente: el poder adquisitivo del dinero.

Los números índices podemos clasificarlos en:

- Números índices simples.

- Números índices compuestos.
  - No ponderados.
  - Ponderados.

3.1. Números índices simples

Estos índices se refieren a un solo artículo o concepto. Son simples relaciones o porcentajes entre dos valores del mismo.

Para la magnitud Y, el índice simple correspondiente al período t, tomando como base el período 0, será:

$$\mathrm{I}_{t/0} = \mathrm{I}_0^t = \frac{Y_t}{Y_0} \times 100$$

Conviene subrayar que el índice simple no es más que el porcentaje que representa Y_t respecto a Y_0. Por lo tanto, carece de dimensión.

Por ejemplo: $\mathrm{I}_0^t = \frac{210}{200} \times 100 = 105$, significa que el valor de la magnitud en el período actual es un 5 % mayor que el valor de la misma magnitud en el período base.

En ocasiones se utilizan los llamados, ÍNDICES EN CADENA, en los que se toma como base el período anterior a aquel en el que se calcula el índice. Su formulación es:

$$\mathrm{I}_{t/t-1} = \mathrm{I}_{t-1}^t = \frac{Y_t}{Y_{t-1}} \times 100$$

Ejemplos:

Las siguientes series corresponden al precio de un artículo en distintos períodos, y vamos a calcular las series de índices simples y de índices encadenados:

t	Y_t	$\mathrm{I}_{t/0}$	$\mathrm{I}_{t/t-1}$
0	125	100	–
1	140	112	112
2	154	123.2	110
3	177.1	141.68	115

t	Y_t	$\mathrm{I}_{t/0}$	$\mathrm{I}_{t/t-1}$
0	140	100	–
1	156.8	112	112
2	180.32	128.8	115
3	216.384	154.56	120

3.2. Números índices compuestos no ponderados

Los índices compuestos son aquellos que hacen referencia a varios artículos o magnitudes. Se trata por tanto de establecer un indicador de la variación experimentada por la característica en estudio, correspondiente al «grupo de artículos o conceptos» contemplándolo como un solo ente.

Supongamos que queremos analizar, por ejemplo, la variación de los precios de un conjunto de magnitudes. En general, para N artículos, la información se puede representar en una tabla de doble entrada, de la siguiente forma:

Magnitudes	1	2	...	N
Período base	Y_{01}	Y_{02}	...	Y_{0N}
Período actual	Y_{t1}	Y_{t2}	...	Y_{tN}
Índices simples	I_1	I_2	...	I_N

El problema consiste en sintetizar la información de la tabla para obtener un indicador que nos ponga de relieve la variación existente entre los precios de los N artículos en el período actual respecto al período base de forma conjunta.

Un criterio para resolver dicho problema es el de utilizar promedios de los números índices simples.

Índice de la media aritmética:

$$\mathrm{I}_{t/0} = \frac{\sum \mathrm{I}_i}{N}$$

Otro criterio para resolver el problema, consiste en calcular un índice simple entre las sumas de los valores de las magnitudes, es decir:

Índice de la media agregativa:

$$\mathrm{I}_{t/0} = \frac{\sum Y_{ti}}{\sum Y_{0i}} \times 100$$

Se debe hacer notar que este índice sólo tiene sentido cuando las magnitudes están medidas en las mismas unidades (no se pueden sumar kg y hl, por ejemplo).

3.3. Números índices compuestos ponderados

Los índices compuestos sin ponderar tienen varios inconvenientes, entre los que destacan los siguientes:

- Al no ponderar los conceptos o magnitudes que intervienen en el índice, esto supone que se otorga la misma importancia a todos ellos.

  La elaboración de un índice debe estar de acuerdo con la finalidad que se persiga, razón por la cual cada magnitud debe venir afectada de un «peso o ponderación» que esté en relación con la importancia que dicha magnitud tiene dentro de todo el conjunto. Así por ejemplo, si se quiere obtener un índice del coste de la vida, el precio de la canela no puede tener la misma importancia que el precio del pan.

- Los artículos pueden medirse en unidades heterogéneas, por lo que no son comparables.

Todo ello ha dado como resultado que los índices sin ponderar tengan un empleo muy limitado, a la vez que da pie a la creación de los índices ponderados.

Por todo esto, en muchas ocasiones, es necesario asignar a cada magnitud simple, y por lo tanto a sus índices, unas ponderaciones que reflejen su peso relativo dentro del conjunto en el que se consideran.

Supongamos que las diferentes ponderaciones asignadas son: $w_1, .., w_i, .., w_N$, de esta forma obtendremos los siguientes índices:

Índice de la media aritmética ponderada:

$$\mathrm{I}_{t/0} = \frac{\sum \mathrm{I}_i w_i}{\sum w_i}$$

Índice de la media agregativa ponderada:

$$\mathrm{I}_{t/0} = \frac{\sum Y_{ti} w_i}{\sum Y_{0i} w_i} \times 100$$

3.4. Índices de precios, de cantidad y de valor

En Economía, los índices más utilizados son los que se refieren a precios, cantidades y valor.

3.4.1. Índices de precios

Podemos considerar los siguientes índices:

- Compuestos sin ponderar:
  - Índice de Sauerbeck:

$$\mathrm{S}_{t/0} = \frac{\sum \frac{P_{ti}}{P_{0i}}}{N} \times 100$$

  - Índice de Bradstreet y Dûtot:

$$\mathrm{BD}_{t/0} = \frac{\sum P_{ti}}{\sum P_{0i}} \times 100$$

- Compuestos ponderados:

  En los índices de precios que se elaboran más frecuentemente, se utilizan como ponderaciones las alternativas siguientes:

  1. $p_{0i}q_{0i}$: representa el valor de las transacciones (precio por cantidad) realizadas para dicho artículo en el periodo base.
  2. $p_{0i}q_{ti}$: (valor ficticio), representa el valor de las transacciones realizadas para dicho artículo en el periodo actual con precios del período base.

Utilizando como ponderación la alternativa 1, y el índice de la media aritmética ponderada de los índices simples, se obtiene el ÍNDICE DE LASPEYRES:

$$\mathrm{L}^p_{t/0} = \frac{\sum \frac{p_{ti}}{p_{0i}} p_{0i} q_{0i}}{\sum p_{0i} q_{0i}} \times 100 = \frac{\sum p_{ti} q_{0i}}{\sum p_{0i} q_{0i}} \times 100$$

Podemos observar que simplificando el índice de Laspeyres también lo podemos definir como una media agregativa ponderada de los precios, usando como ponderación las cantidades en el período base.

La elaboración de un índice de Laspeyres tiene la ventaja, y por ello es el que más se utiliza, de que las ponderaciones del período base se mantienen fijas para todos los períodos. Sin embargo, presenta el inconveniente de que pierde representatividad a medida que nos alejamos del período base.

Cuando se utiliza la alternativa 2 como ponderación en una media aritmética de índices simples, se obtiene el ÍNDICE DE PAASCHE:

$$\mathrm{P}^p_{t/0} = \frac{\sum \frac{p_{ti}}{p_{0i}} p_{0i} q_{ti}}{\sum p_{0i} q_{ti}} \times 100 = \frac{\sum p_{ti} q_{ti}}{\sum p_{0i} q_{ti}} \times 100$$

Este índice también se puede ver como una media agregativa de los precios, siendo las ponderaciones las cantidades en el momento actual.

En este índice las ponderaciones ($p_{0i}q_{ti}$) son variables. Concretamente, para su elaboración se requiere información de los precios y cantidades en cada período, a diferencia del de Laspeyres, para cuya elaboración únicamente se precisa información sobre las cantidades del período base, aparte, claro está, de los datos sobre precios de cada periodo.

El índice de Paasche también pierde representatividad, aunque en menor medida que el índice de Laspeyres, a medida que el año con el que se efectúa la comparación, está más alejado del año base.

Otro índice ponderado, aunque menos utilizado es el:

ÍNDICE DE FISHER: es la media geométrica de los índices de Laspeyres y Paasche, con lo cual, su valor estará acotado por el valor que tienen ambos índices.

$$\mathrm{F}^p_{t/0} = \sqrt{\mathrm{L}^p_{t/0}\mathrm{P}^p_{t/0}} = \sqrt{\frac{\sum p_{ti} q_{0i}}{\sum p_{0i} q_{0i}} \times \frac{\sum p_{ti} q_{ti}}{\sum p_{0i} q_{ti}}} \times 100$$

3.4.2. Índices de cantidad

Son los que tratan de medir la evolución relativa de una magnitud económica (producción, consumo, etc...) en términos reales, es decir, sin recoger el efecto que sobre ella pueda haber tenido la variación de precios.

Solo nos vamos a fijar en las formulaciones de números índices compuestos ponderados, ya que únicamente se suelen utilizar estos.

Al igual que en el caso de los números índices de precios, los índices cuánticos más utilizados son los de Laspeyres y Paasche.

- Índice de Laspeyres:

$$\mathrm{L}^q_{t/0} = \frac{\sum \frac{q_{ti}}{q_{0i}} p_{0i} q_{0i}}{\sum p_{0i} q_{0i}} \times 100 = \frac{\sum q_{ti} p_{0i}}{\sum q_{0i} p_{0i}} \times 100$$

- Índice de Paasche:

$$\mathrm{P}^q_{t/0} = \frac{\sum \frac{q_{ti}}{q_{0i}} q_{0i} p_{ti}}{\sum q_{0i} p_{ti}} \times 100 = \frac{\sum q_{ti} p_{ti}}{\sum q_{0i} p_{ti}} \times 100$$

El índice de cantidad de Fisher será, en este caso:

Índice de Fisher:

$$\mathrm{F}^q_{t/0} = \sqrt{\mathrm{L}^q_{t/0} \mathrm{P}^q_{t/0}}$$

3.4.3. Índices de valor

El valor de un conjunto de mercancías (producidas, consumidas, exportadas, etc...) o gasto, para dos períodos de tiempo, el actual y el base, vendrá dado, respectivamente, por las siguientes expresiones:

$$V_t = \sum V_{ti} = \sum p_{ti} q_{ti}$$
$$V_0 = \sum V_{0i} = \sum p_{0i} q_{0i}$$

El cociente entre ambos agregados es:

$$\mathrm{I}^V_{t/0} = \frac{V_t}{V_0} \times 100 = \frac{\sum V_{ti}}{\sum V_{0i}} \times 100 = \frac{\sum p_{ti} q_{ti}}{\sum p_{0i} q_{0i}} \times 100$$

y se denomina **índice de valor** agregado de la producción o del gasto en consumo.

Resulta evidente que en un índice de valor se reflejan conjuntamente las variaciones de los precios y las cantidades, ya que las variaciones entre los valores no son sino el efecto conjunto de las variaciones de las cantidades y de sus precios respectivos entre ambos períodos.

Si para un cierto artículo se verifica que su valor es igual al precio por la cantidad, $v = pq$, parece lógico que exijamos que para un grupo de artículos se cumpla la misma ecuación y por lo tanto, para los índices se debería exigir que $\mathrm{I}^V = \mathrm{I}^P \mathrm{I}^Q$.

Es fácil demostrar que esto no siempre es cierto. En particular, para los índices que hemos definido:

1. $\mathrm{L}^p\mathrm{L}^q \neq \mathrm{I}^V$
2. $\mathrm{P}^p\mathrm{P}^q \neq \mathrm{I}^V$
3. $\mathrm{F}^p\mathrm{F}^q = \mathrm{I}^V$

También es evidente que (prescindiendo del 100):

$$\mathrm{I}^V_{t/0} = \frac{V_t}{V_0} = \frac{\sum V_{ti}}{\sum V_{0i}} = \frac{\sum p_{ti}q_{ti}}{\sum p_{0i}q_{0i}} = \begin{cases} \frac{\sum p_{ti}q_{ti}}{\sum p_{0i}q_{ti}} \times \frac{\sum p_{0i}q_{ti}}{\sum p_{0i}q_{0i}} = \mathrm{P}^p_{t/0}\mathrm{L}^q_{t/0} \\ \frac{\sum p_{ti}q_{ti}}{\sum p_{ti}q_{0i}} \times \frac{\sum p_{ti}q_{0i}}{\sum p_{0i}q_{0i}} = \mathrm{P}^q_{t/0}\mathrm{L}^p_{t/0} \end{cases}$$

Es decir que, el índice de valor puede expresarse mediante un producto de índices de Laspeyres (precios o cantidades) y de Paasche (cantidades o precios) respectivamente.

3.5. Propiedades de los números índices

Irving Fisher intentó establecer una sistematización de los números índices, y propuso una serie de criterios o propiedades para examinar las diferentes fórmulas, de tal forma que un índice será más ventajoso cuantas más propiedades cumpla (esto no quiere decir que si no las satisface la fórmula deba eliminarse).

Estas propiedades parten del siguiente principio: *Lo que es cierto para un producto, y por lo tanto para un índice simple, debería ser cierto para un conjunto de ellos, y en consecuencia, para el número índice compuesto que los representa.*

En estas propiedades se prescinde de multiplicar por 100.

- **Identidad**: se dice que un índice cumple el criterio de identidad, si el índice del período t, respecto al período t como base, es la unidad.

  $$\mathrm{I}_{t/t} = 1$$

  La cumplen todos.

- **Inversión**: se dice que un índice cumple el criterio de inversión si el índice respecto a t con base t', multiplicado por el índice respecto a t' con base t es la unidad.

  $$\mathrm{I}_{t/t'} \times \mathrm{I}_{t'/t} = 1$$

  La verifican: BD; F. No la verifican: S; L; P.

- **Reversibilidad de los factores**: se dice que un índice cumple el criterio de reversibilidad de los factores si el índice de precios por el índice de cantidad es el índice de valor:

  $$\mathrm{I}^p_{t/0} \times \mathrm{I}^q_{t/0} = \mathrm{I}^V_{t/0}$$

  La verifica: F. No la verifican: S; BD; L; P.

- **Transitiva o circular** (es una generalización de la propiedad de inversión):

$$\mathrm{I}_{t'/t} = \mathrm{I}_{t'/t'-1} \times \mathrm{I}_{t'-1/t'-2} \times \cdots \times \mathrm{I}_{t+2/t+1} \times \mathrm{I}_{t+1/t}$$

  La verifica: BD. No la verifica: F.

3.6. Pasos para el cálculo de los números índices

Al elaborar un índice compuesto, hay que realizar una serie de pasos, entre los que destacan los siguientes (nos referiremos al cálculo del índice de precios al consumo):

1. **Selección de variables**: El primer problema que se plantea es el de seleccionar «qué variables» entrarán en el mismo y definirlas perfectamente. Ej.: Cacao: Nesquik de 800 gr.

   Puesto que el índice es un resumen del grupo o conjunto al que se refiere, se deben elegir los artículos o conceptos más relevantes dentro del grupo (pan vs canela).

2. **Selección de los lugares y tiempos de observación**: Una vez definidos los conceptos que forman el grupo, se procederá a conseguir las observaciones. Estas observaciones consisten en la obtención de los valores numéricos correspondientes a los precios y/o cantidades de los artículos seleccionados.

   Es necesario que las observaciones se lleven a cabo siempre en los mismos lugares y referentes al mismo tipo o variedad de artículo. También debe especificarse el instante en el que se lleva a cabo la observación (no promociones, aunque ahora entran las rebajas), o bien un intervalo de tiempo al que se referirán las distintas tomas de datos.

   Los lugares y tiempos de observación se deben seleccionar en función de la importancia del concepto dentro del grupo.

3. **Selección de la base**: Puesto que el tiempo base es el término de referencia o comparación, se debe elegir de forma que sea un tiempo o época «normal».

   Si, por ejemplo, vamos a elaborar un índice de producción agrícola, no debemos tomar como año base un año de cosechas excepcionales, ya que el resto de los años, al compararlos, nos arrojarán datos infravalorados. Por ello, en ciertos tipos de índices de producción, en los que, por las características de los artículos, se presentan grandes fluctuaciones, se recomienda tomar como base un promedio de varios años, para eliminar así dichas fluctuaciones.

4. **Selección de fórmulas y ponderaciones**: Se debe tener en cuenta que las distintas fórmulas guardan una estrecha relación con las ponderaciones y con el coste en que se incurre para elaborar el índice.

   En el caso de no disponer de las ponderaciones, solo podemos elaborar índices sin ponderar (Sauerbeck o Bradstreet). En el caso de que se conozcan las ponderaciones, la elección depende de los medios de que se disponga. El más costoso es el de Fisher, seguido del de Paasche y el de Laspeyres. Este último solo necesita conocer las ponderaciones del año base, por lo que es el más económico.

5. **Renovación del índice**: A medida que el tiempo transcurre (y nos alejamos de la base) tienen lugar cambios en el comportamiento de las variables, de modo que el conjunto de variables que se ha seleccionado para la elaboración del índice puede haber dejado de ser representativo. Por otro lado, las ponderaciones pueden no ajustarse tampoco a la estructura del momento.

   Ej.: Hace unos cincuenta años, el pollo solo se consumía los domingos (y solo algunos privilegiados) y ahora se consume todos los días. Por otra parte, en los hogares se consumía carbón y por supuesto no había ordenadores.

   Por ello, conviene «renovar» el índice. Esto lleva consigo *comenzar todo el proceso desde el principio: elegir las variables, nueva base, ponderaciones, ...*

6. **Empalme o enlace de índices nuevos con los antiguos**: Una vez que se ha llevado a cabo la renovación del índice, nos encontraremos series de índices, cada una de las cuales abarcará períodos distintos. La serie nueva comenzará naturalmente por 100, y habrá entonces una «rotura» de datos.

   En muchas ocasiones necesitaremos una sola serie de índices que nos permita ver la evolución temporal del fenómeno, para ello debemos enlazar las dos series de índices. El proceso es muy sencillo, veámoslo sobre un ejemplo:

   Supongamos que tenemos los índices de precios al consumo en dos series. La primera, con base en 2001, abarca los años 2003 al 2006, y la segunda, con base en 2006, abarca los años 2006 a 2009, es decir que se dispone de los siguientes datos:

	IPC base 2001	IPC base 2006
2003	106.68	
2004	109.93	
2005	113.63	
2006	117.62	100
2007		102.80
2008		107.00
2009		106.70

   El enlace se puede hacer con base en 2001 o con base en 2006.

   Suponiendo que queramos obtener toda la serie con base en este último año, estableceremos la siguiente regla de tres:

   Si 117.62 equivale a 100 entonces 113.63 equivale a x , por lo tanto:

   $x = 113.63 \times \frac{100}{117.62} = 113.63 \times \frac{1}{1.1762}$

   Luego, para pasar de base 2001 a base 2006, tenemos que dividir los índices con base en 2001 por la constante: 1.1762

   De forma similar se obtiene la serie con base en 2001, sin más que multiplicar los índices con base en 2006 por la constante: 1.1762

	IPC base 2001	IPC base 2006
2003	106.68	**90.70**
2004	109.93	**93.46**
2005	113.63	**96.61**
2006	117.62	100
2007	**120.91**	102.80
2008	**125.85**	107.00
2009	**125.50**	106.70

Hay que hacer constar que, en realidad, **los dos períodos enlazados no son estrictamente comparables**, ya que al renovar la base pueden haber entrado en el conjunto artículos nuevos, habrán desaparecido otros, habrán cambiado las ponderaciones, etc..., pero cuando es necesario enlazar dos series no queda otra solución.

Sin embargo, no hay que olvidar que el índice es solo «un reflejo» de la variación del fenómeno, y no una medida exacta, y por lo tanto, como **indicador** de estas variaciones, puede sernos de utilidad, a pesar de los inconvenientes citados.

7. **Cambio de base**: En muchas ocasiones necesitaremos expresar los índices calculados con base en una época 0, en otra base t'. La diferencia de este caso con el anterior, es que ahora no existe renovación del índice, sino que artificialmente vamos a cambiar de período base, conservando las ponderaciones del periodo base 0, respecto al que se elaboraron los índices.

   Por ejemplo, supongamos que conocemos los índices de precios al consumo, calculados con base en 2001:

	IPC base 2001
2001	100
2002	103.54
2003	106.68
2004	109.93
2005	113.63
2006	117.62

   Si queremos tomar como base el año 2003 (IPC=100) podemos operar igual que antes (resolviendo una regla de tres): Como el valor 106'68 hay que convertirlo en 100, habrá que dividirlo por 1,0668; por lo tanto este es el valor constante por el que habrá que dividir todos los términos de la serie.

	IPC base 2001	**IPC base 2003**
2001	100	**93.74**
2002	103.54	**97.06**
2003	106.68	**100**
2004	109.93	**103.05**
2005	113.63	**106.51**
2006	117.62	**110.25**

   De esta forma se obtienen los valores de la serie con base en 2003.

   En estos casos, para evitar confusiones, se debe especificar, además del año base, el año al que se refieren las ponderaciones.

3.7. La deflación de valores

Ya hemos comentado que para solucionar el problema de la agregación de «bienes y servicios» heterogéneos, se procede a la valoración de los mismos. En economía es general el empleo que se hace de estos valores.

La manera de calcular el valor de un bien, consiste en multiplicar la cantidad (q) por el precio (p); y el precio tiene que venir expresado en unidades monetarias, es decir, en **dinero**.

Pero esta unidad de medida no es fija, y a lo largo del tiempo sufre alteraciones, que se concretan en variaciones del valor del dinero. Lo normal es que los precios se eleven a medida que transcurre el tiempo, proceso conocido con el nombre de **inflación**. La inflación, por lo tanto, origina una pérdida en el valor del dinero (pérdida de poder adquisitivo).

Al **comparar** valores correspondientes a dos épocas distintas, debemos tener en cuenta que deben estar expresadas en **unidades monetarias equivalentes**, es decir, con el mismo poder adquisitivo; sin embargo, al venir referidas a tiempos distintos, vendrán expresadas en unidades monetarias con diferente poder adquisitivo.

Por lo tanto es necesario corregir la pérdida de valor del dinero, para obtener una unidad de medida homogénea, o lo que es lo mismo: expresar los valores en **unidades monetarias con poder adquisitivo constante**.

El procedimiento mediante el cual corregimos la pérdida del valor del dinero se conoce con el nombre de **deflación**.

Entonces, a la hora de comparar magnitudes económicas en valor a lo largo del tiempo, se requiere que estos valores sean homogéneos, lo cual requiere *deflactar la serie de valores corrientes mediante un índice de precios adecuado.* El índice que se utiliza para realizar esta operación recibe el nombre de deflactor.

Para obtener la **serie deflactada** (en la que todos los valores deben estar expresados en las mismas unidades monetarias) hay que **dividir la serie original** en precios corrientes, **por el deflactor correspondiente**; de esta forma, la nueva serie refleja su evolución real en el tiempo, independientemente de las alteraciones monetarias.

No existe un deflactor único, sino que depende de la magnitud que se trate de obtener. Así por ejemplo, si se quiere medir la capacidad real de compra de los consumidores privados en bienes de consumo, habrá que deflactar las rentas monetarias de cada período por el índice deflactor correspondiente, que sería, en este caso, el índice de precios al consumo.

Veamos ahora qué posibilidades ofrecen los índices de Laspeyres y Paasche para deflactar una serie económica agregada.

Sean:

$V_t = \sum V_{ti} = \sum p_{ti}q_{ti}$, valor agregado a precios corrientes del período actual.

$V_0 = \sum V_{0i} = \sum p_{0i}q_{0i}$, valor agregado a precios corrientes del período base.

Al dividir V_t por el índice de precios de Laspeyres, resulta la siguiente expresión:

$$\frac{V_t}{\mathrm{L}^p} = \frac{\sum p_{ti}q_{ti}}{\frac{\sum p_{ti}q_{0i}}{\sum p_{0i}q_{0i}}} = V_0 \times \mathrm{P}^q$$

Mientras que si deflactamos V_t mediante un índice de precios de Paasche, tenemos que:

$$\frac{V_t}{\mathrm{P}^p} = \frac{\sum p_{ti}q_{ti}}{\frac{\sum p_{ti}q_{ti}}{\sum p_{0i}q_{ti}}} = \sum p_{0i}q_{ti}$$

Así pues, al deflactar V_t mediante un índice de precios de *Laspeyres*, se obtiene como resultado una *proyección temporal del valor inicial, V_0, a través de un índice cuántico de Paasche*.

Por el contrario, al deflactar con un índice de precios de **Paasche** se obtiene la **valoración de la producción actual a precios del período base**. Por lo tanto, el índice de Paasche es el deflactor idóneo, ya que permite pasar de valores monetarios corrientes a valores expresados en precios (los del período base) constantes. Si utilizamos como deflactor cualquier otro índice, no obtenemos valores a precios constantes.

A pesar de que el índice de Paasche es el más adecuado, en la práctica se utiliza en muchas ocasiones el índice de Laspeyres por ser el único disponible, ya que el primero exige para su elaboración una información que habitualmente no está disponible.

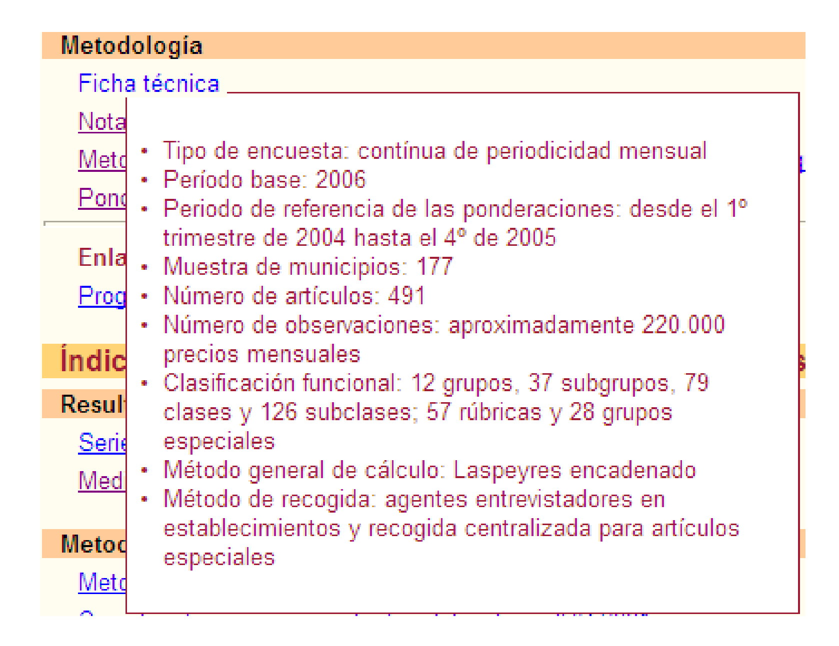
Por otra parte, aunque en la expresión anterior hemos considerado valores que se puedan descomponer en suma de productos de precios por cantidades, se puede presentar el problema de deflactar una magnitud macroeconómica que no permita tal descomposición. Por ejemplo, puede interesarnos expresar, en términos constantes, la renta personal disponible, que es una magnitud que tiene un carácter estrictamente monetario.

Para deflactar una magnitud de este tipo, debe determinarse previamente el objetivo perseguido con tal operación, para después proceder a la elección del deflactor más adecuado para tal efecto. Así, si lo que queremos es expresar la renta personal disponible en términos de poder adquisitivo en bienes de consumo, el deflactor a elegir será el índice de precios de consumo.

3.8. Índice de precios de consumo

El Índice de Precios de Consumo (IPC) es una medida estadística de la evolución del conjunto de precios de los bienes y servicios que consume la población residente en viviendas familiares en España.

La ficha técnica del IPC actual es la que se muestra a continuación:



Para saber más sobre la metodología general del IPC calculado en España, se recomienda leer el documento Metodología que se encuentra en la página del INE (**www.ine.es**):

http://www.ine.es/daco/daco43/metoipc06.pdf

3.9. Ejemplos resueltos

Veamos algunas de las muchas cuestiones que se pueden resolver con números índices.

1. En la siguiente tabla se muestran los salarios medios mensuales, en euros, de cierta categoría de empleados, así como los índices de precios de consumo en el mismo período:

Año	Salario	$IPC_{t/06}$
2004	2043	93.5
2005	2125	96.6
2006	2252	100
2007	2393	102.8
2008	2513	107
2009	2561	106.7

   a) Obtén los índices que miden la variación de los salarios de cada año respecto al año anterior.

   b) Obtén los índices que miden la variación de los precios de cada año respecto al año 2004.

   c) Expresa la serie de los salarios anuales en unidades monetarias constantes de 2004.

   Solución.

a) Obtén los índices que miden la variación de los salarios de cada año respecto al año anterior.

Tenemos que calcular los índices encadenados:

$$\mathrm{IS}_{t/t-1} = \frac{S(t)}{S(t-1)} \times 100$$

El primero que podemos calcular es:

$$\mathrm{IS}_{05/04} = \frac{2125}{2043} \times 100 = 104.01$$

análogamente:

$$\mathrm{IS}_{06/05} = \frac{2252}{2125} \times 100 = 105.98$$

y así con todos (al final del ejercicio está la tabla completa).

b) Obtén los índices que miden la variación de los precios de cada año respecto al año 2004.

En este caso tenemos que hacer un cambio de base en el índice.Ahora el año base es 2004, por lo que tenemos que convertir el 93.5 en un 100, para lo cual nos basta con multiplicar la serie del IPC por $\frac{100}{93.5}$

$$\mathrm{IPC}_{t/04} = \mathrm{IPC}_{t/04}\frac{100}{93.5}$$

Los correspondientes valores del IPC con base en 2004 están en la tabla, al final del ejercicio.

c) Expresa la serie de los salarios anuales en unidades monetarias constantes de 2004.

Los salarios están expresados en unidades monetarias corrientes de cada año y queremos expresarlos en **términos constantes del año 2004**.

Para ello debemos eliminar el efecto de la inflación (variación de los precios de cada año respecto al año 2004), por lo tanto, debemos deflactar, dividiendo los salarios en u.m. corrientes, por el índice de precios de cada año **con base en el año 2004**.

Así, en el año 2004 el salario medio mensual fue de 2043.

En el año 2005, fue de $\frac{2125}{1.0332} = 2056$ (euros de 2004)

y así con todos los demás.

Año	$S(t)$	$\mathrm{IPC}_{t/06}$	$\mathrm{IS}_{t/t-1}$	$\mathrm{IPC}_{t/04}$	S(ctes de 2004)
2004	2043	93.5	—	100.00	2043.00
2005	2125	96.6	104.01	103.32	2056.81
2006	2252	100	105.98	106.95	2105.62
2007	2393	102.8	106.26	109.95	2176.51
2008	2513	107	105.01	114.44	2195.94
2009	2561	106.7	101.91	114.12	2244.18

2. La familia Pérez ha ido registrando los ingresos del hogar en los últimos 5 años y dispone también de los datos del IPC con base en el año 2006.

Año	Ingreso	$IPC_{t/06}$
2004	28602	93.5
2005	29750	96.6
2006	31528	100
2007	33502	102.8
2008		107

a) Si los ingresos del hogar aumentaron en 2008 un 2 % respecto al año anterior ¿cuál fue el ingreso del hogar ese año?

b) Viendo la evolución del IPC ese año, la familia dice haber perdido poder adquisitivo y exige al Gobierno una paga para compensar esta pérdida ¿de cuánto debería ser esa paga?

c) Calcula el ingreso del hogar en el año 2007 en términos constantes del año 2004.

Solución.

a) Si los ingresos del hogar aumentaron en 2008 un 2 % respecto al año anterior ¿cuál fue el ingreso del hogar ese año?

El ingreso de 2008 fue el ingreso de 2007 más un 2 % de dicho ingreso:

$$\text{Ingreso}(08) = \text{Ingreso}(07) + \frac{2}{100}\text{Ingreso}(07) = 1.02\ \text{Ingreso}(07)$$

$$\text{Ingreso}(08) = 1.02 \times 33502 = 34172.04 \text{ euros}$$

b) Viendo la evolución del IPC ese año, la familia dice haber perdido poder adquisitivo y exige al Gobierno una paga para compensar esta pérdida ¿de cuánto debería ser esa paga?

Para no perder poder adquisitivo la variación de los ingresos debería haber sido igual a la del IPC.

Calculamos la variación de los precios de 2008 respecto a 2007:

$$\text{IPC}_{08/07} = \frac{\text{IPC}_{08/06}}{\text{IPC}_{07/06}} \times 100 = \frac{107}{102.8} \times 100 = 104.0856$$

Los precios subieron un 4.0856 %, por lo que los ingresos deberían haber subido en la misma proporción. Es decir que para no perder poder adquisitivo el ingreso debería haber sido de:

$$\text{Ingreso correcto}(2008) = 33502 \times 1.040856 = 34870.76 \text{ euros}$$

Por lo tanto, la familia debería solicitar una paga por la diferencia:

$$\text{Paga} = 34870.76 - 34172.04 = 698.72 \text{ euros}$$

c) Calcula el ingreso del hogar en el año 2007 en términos constantes del año 2004.

Para calcular el ingreso del hogar en el año 2007 en términos constantes del año 2004, debemos expresar dicha cantidad eliminando el efecto de la inflación en ese período.

Calculamos la variación de los precios para 2007 respecto a 2004:

$$\text{IPC}_{07/04} = \frac{\text{IPC}_{07/06}}{\text{IPC}_{04/06}} \times 100 = \frac{102.8}{93.5} \times 100 = 109.9465$$

Y ahora deflactamos el ingreso de 2007:

$$\text{Ingreso}_{07}(\text{en términos ctes de 2004}) = \frac{\text{Ingreso}_{07}}{\text{IPC}_{07/04}}$$

$$\text{Ingreso}_{07}(\text{en términos ctes de 2004}) = \frac{33502}{1.099465} = 30471.18 \text{ euros de 2004}$$

Tema 4

La curva Normal

Se dice que algo es normal, cuando se encuentra en su estado natural, cuando sirve de norma o regla, o cuando por su naturaleza, forma o magnitud se ajusta a ciertas normas fijadas de antemano.

Y ¿qué tiene que ver esto con la Estadística?, pues mucho más de lo que parece.

Cuando estudiamos una característica de una población, nos interesa saber si los valores observados son normales, es decir, si el comportamiento de nuestra variable, en la población analizada, es normal, es el esperado o el que cabría esperar, o si, por el contrario, la variable presenta un comportamiento anómalo.

Si pensamos en la altura o el peso de los hombres adultos de una determinada población, podemos observar que hay unos determinados valores que nos pueden parecer normales (175 cm, 80 kg), y que nos lo parecen así porque son los más habituales, los que aparecen con mayor frecuencia, mientras que los valores alejados de estos, tanto por exceso como por defecto ya no se consideran normales (225 cm, 40 kg) y si aparecen lo hacen con una frecuencia muy pequeña. En general, lo normal, se encuentra cerca del valor medio y es lo más frecuente.

Esta idea la plasmó Gauss en una curva llamada **curva Normal**, cuya formulación matemática es la siguiente:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

donde:

$f(x)$ es la frecuencia de un determinado valor

x es un valor cualquiera de la variable

μ es la media de la distribución

σ es la desviación típica de la distribución

π es la constante: 3.14159...

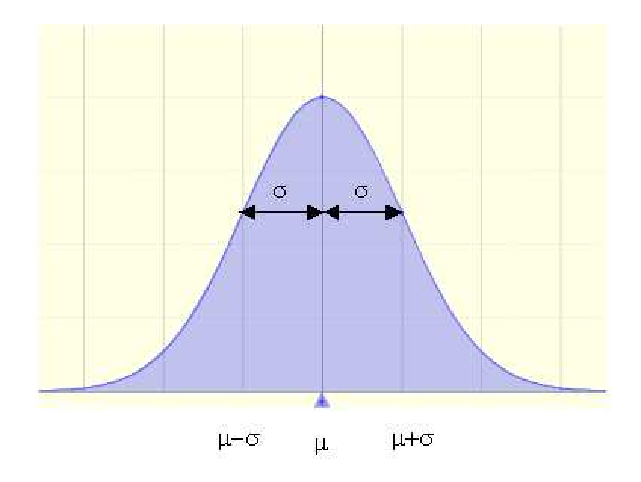
e es la constante: 2.71828...

Los valores de μ y de σ constituyen los parámetros de la curva Normal, que denotaremos como $N(\mu, \sigma)$ (Normal de media μ y desviación típica σ).

Está claro que este no es más que un modelo teórico y que ningún fenómeno de la naturaleza se va a ajustar exactamente a este modelo, pero sí que hay muchos fenómenos cuyo comportamiento se acercará mucho.

Más adelante (en las prácticas con ordenador) veremos cómo determinar si nuestra variable tiene un comportamiento parecido al de una Normal, pero por ahora vamos a estudiar esta función y a descubrir algunas de sus propiedades.

4.1. Propiedades de la curva Normal



A simple vista se pueden observar varias características:

1. Tiene forma de campana.

2. Es simétrica respecto al parámetro μ.

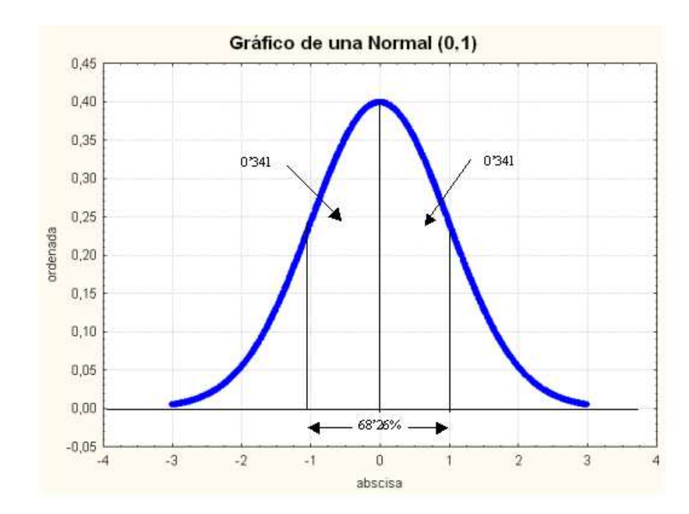
3. La media, la mediana y la moda coinciden con el valor μ.

4. Los puntos de inflexión de la curva corresponden a las abscisas $\mu \pm \sigma$.

   Además esta curva tiene otras propiedades que, aunque se intuyen, no se ven a simple vista y que vamos a comentar:

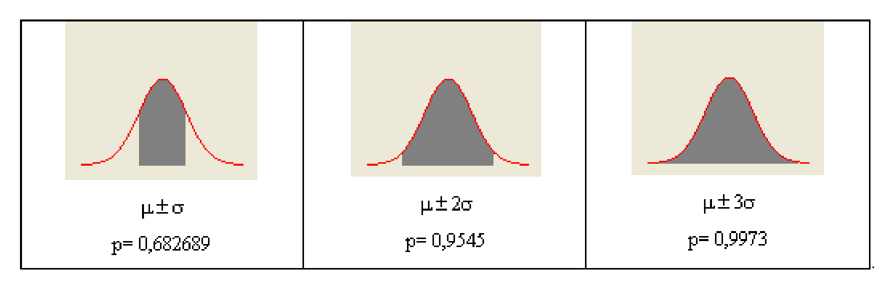
5. El área total bajo la curva vale 1.

6. Existe una relación muy interesante entre la media y la desviación típica: «*la proporción de datos que se encuentran entre la media y la media más una desviación típica es de 0.3413*» (aproximadamente un tercio).



Esto significa que en el intervalo de 2 desviaciones típicas en torno a la media, se concentra una proporción de observaciones de 0.6826, o lo que es lo mismo, el 68.26 % de las observaciones (algo más de dos tercios).

En general, casi el 100 % de las observaciones se encuentran a menos de 3 desviaciones típicas de la media.



Nota: estamos identificando la proporción de datos en un intervalo con el área bajo la curva en dicho intervalo.

4.2. Valores tipificados

Una utilidad de la información anterior es la siguiente:

Supongamos que conocemos la media (μ=6.5) y la desviación típica (σ=1.7) de las calificaciones de los alumnos de esta asignatura en el primer trabajo del curso, y que sabemos que dichas calificaciones se distribuyen normalmente (es decir, siguen una distribución Normal). Entonces, podemos afirmar que el 68.27 % de los estudiantes de la clase, tiene una calificación entre 6.5-1.7 (=4.8) y 6.5+1.7 (=8.2).

Otra situación que se puede plantear es la siguiente: un compañero nos dice que ha sacado un 7.5 en los dos trabajos que hay que realizar durante el curso, ¿cómo interpretamos estas puntuaciones?

Directamente no lo podemos interpretar, pero, suponiendo que las calificaciones del segundo trabajo sigan también una distribución Normal, y si conocemos la media (μ=8) y

la desviación típica (σ=1.5) de las notas del grupo, podemos comparar ambas calificaciones y determinar cuál es su posición en el grupo en ambos casos.

Por un lado, podemos calcular su desviación respecto a la media del grupo:

En el primer trabajo: 7.5-6.5=1, su calificación está 1 punto por encima de la media del grupo, mientras que en el segundo trabajo: 7.5-8=-0.5, lo que significa que su nota está medio punto por debajo de la nota media del grupo.

Sin embargo, como ya sabemos, es importante conocer lo próximos o alejados que se encuentran los valores de la media, por lo que, si dividimos estas desviaciones por la desviación típica (es decir, utilizamos la desviación típica como unidad de medida de la dispersión), obtendremos unos valores, llamados valores tipificados, que corresponden a distribuciones del mismo tipo (estos valores corresponden a una escala que tiene el 99.73 % de sus valores entre -3 y 3).

$$z_1 = \frac{7.5 - 6.5}{1.7} = 0.59 \text{ , mientras que en el segundo trabajo: } z_2 = \frac{7.5 - 8}{1.5} = -0.33$$

Ahora, ambos valores corresponden a la misma escala y son comparables. Como vemos, las dos notas, aunque numéricamente son iguales, no representan lo mismo.

La nota del primer trabajo está 0.59 veces la desviación típica, por encima de la nota media del grupo, mientras que la nota del segundo trabajo está 0.33 veces la desviación típica por debajo de la nota media del grupo.

Como ambos valores están en la misma escala, podemos afirmar que es mucho mejor nota la del primer trabajo, que la nota del segundo trabajo, con relación a las notas del grupo.

Los valores tipificados nos permiten comparar tanto los valores de un mismo sujeto para distintas variables (que pueden estar medidas en distintas escalas), como los valores de distintos sujetos para la misma variable.

Características de los valores tipificados

1. Los valores tipificados son una mera transformación lineal de los valores observados y por lo tanto son equivalentes.

   Esto significa que la forma de la distribución de los valores tipificados es la misma que la de los valores originales.

2. La media de los valores tipificados es siempre cero.

   Esto es consecuencia directa de la propiedad de la media que dice que la suma de las desviaciones respecto a la media es cero.

$$Z = \frac{X - \bar{x}}{s'} \Rightarrow \bar{z} = \frac{1}{N} \sum_{i=1}^{k} z_i n_i = \frac{1}{N} \sum_{i=1}^{k} \left(\frac{x_i - \bar{x}}{s'} \right) n_i = \frac{1}{N} \frac{1}{s'} \sum_{i=1}^{k} (x_i - \bar{x}) n_i = \frac{1}{N} \frac{1}{s'} 0 = 0$$

3. La desviación típica de los valores tipificados es siempre 1.

   Esto se debe a las propiedades de la varianza (y por lo tanto de la desviación típica) frente a transformaciones lineales:

$$Z = \frac{X - \bar{x}}{s'} \text{ donde } s' = \mathrm{S}'_X \text{ y sabemos que } \mathrm{S}'^2_Z = \frac{1}{N}\sum_{i=1}^{k}(z_i - \bar{z})^2 n_i$$

$$\text{Como } z_i - \bar{z} = \frac{x_i - \bar{x}}{s'}, \text{ entonces:}$$

$$\mathrm{S}'^2_Z = \frac{1}{N}\sum_{i=1}^{k}(z_i - \bar{z})^2 n_i = \frac{1}{N}\sum_{i=1}^{k}\left(\frac{x_i - \bar{x}}{s'}\right)^2 n_i = \frac{1}{s'^2}\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i = \frac{s'^2}{s'^2} = 1$$

4.3. Proporciones de la curva Normal

Suponiendo que nuestra distribución se ajusta a una curva Normal y que conocemos la media y la desviación típica, podemos averiguar la proporción de datos que cumplen determinados criterios.

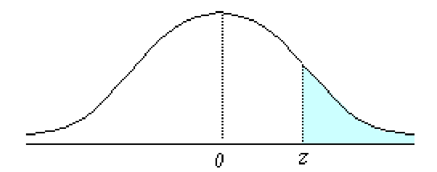
Para ello se utilizan los valores (que están tabulados) de la Normal de media 0 y desviación típica 1. Tabla de la $N(0,1)$:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014

Para valores mayores:

z	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	0.00135	$0.0^{3}968$	$0.0^{3}687$	$0.0^{3}483$	$0.0^{3}337$	$0.0^{3}233$	$0.0^{3}159$	$0.0^{3}108$	$0.0^{4}723$	$0.0^{4}481$
4	$0.0^{4}317$	$0.0^{4}207$	$0.0^{4}133$	$0.0^{5}854$	$0.0^{5}541$	$0.0^{5}340$	$0.0^{5}211$	$0.0^{5}130$	$0.0^{6}793$	$0.0^{6}479$
5	$0.0^{6}287$	$0.0^{6}170$	$0.0^{7}996$	$0.0^{7}579$	$0.0^{7}333$	$0.0^{7}190$	$0.0^{7}107$	$0.0^{8}599$	$0.0^{8}332$	$0.0^{8}182$
6	$0.0^{9}987$	$0.0^{9}530$	$0.0^{9}282$	$0.0^{9}149$	$0.0^{10}777$	$0.0^{10}402$	$0.0^{10}206$	$0.0^{10}104$	$0.0^{11}523$	$0.0^{11}260$
7	$0.0^{11}128$	$0.0^{12}624$	$0.0^{12}301$	$0.0^{12}144$	$0.0^{13}682$	$0.0^{13}320$	$0.0^{13}149$	$0.0^{14}688$	$0.0^{14}311$	$0.0^{14}133$

Esta tabla representa la proporción de observaciones que se encuentran «a la derecha» de un determinado valor z, correspondiente a una variable que se distribuye según una Normal de media 0 y desviación típica 1: $N(0, 1)$.



4.3.1. ¿Cómo se utiliza la tabla?

En primer lugar, sabemos que la curva es simétrica, por lo que la mitad de las observaciones (0.5 o el 50 %), se encuentran en cada una de las dos mitades. Por eso solo se utiliza la parte de la derecha, ya que haciendo un cálculo muy sencillo se pueden obtener las proporciones correspondientes para los valores negativos.

¿Cómo se leen los valores de la tabla?

En general se trabaja con valores típicos con dos decimales. La parte entera y el primer decimal están en la columna de la izquierda de la tabla, y el segundo decimal en la primera fila.

De este modo, para buscar la proporción de observaciones con un valor típico mayor que 0.59, tenemos que buscar la intersección entre la fila del 0.5 y la columna del 0.09:

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148

Es decir que dicha proporción es de 0.2776, o dicho de otra forma, un 27.76 % de las observaciones tienen una puntuación típica mayor que 0.59.

Si recordamos el ejemplo de las calificaciones en los trabajos, estaríamos diciendo que solo el 27.76 % de los compañeros de clase tienen una puntuación mejor en el primer trabajo. Esto también se puede interpretar diciendo que el 72.24 % de sus compañeros tienen una calificación inferior.

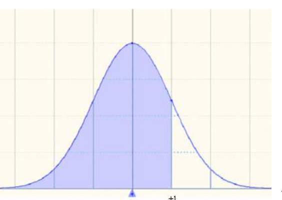
Como las distribuciones de los datos originales y de los valores típicos correspondientes son equivalentes, podemos afirmar que en el primer trabajo: 7.5 es una nota que no es superada por el 72.24 % de los alumnos del curso; o bien, 7.5 es una nota que solo es superada por el 27.76 % de los alumnos del curso.

RECUERDA: para utilizar la tabla de la Normal, los valores deben corresponder a una distribución Normal de media 0 y desviación típica 1, por lo tanto, para poder calcular todas estas proporciones, debemos tipificar previamente.

4.3.2. Cálculos en distintas situaciones

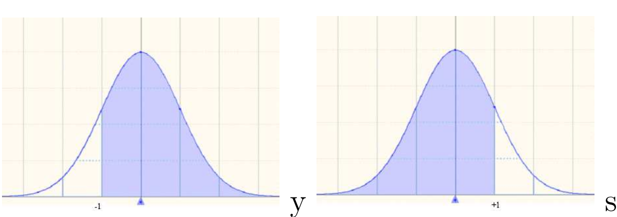
La tabla solo nos da la proporción de datos por encima de un determinado valor positivo, así que si queremos calcular alguna otra proporción tendremos que hacer algunos cálculos para obtenerla.

a) Proporción de datos por debajo de un determinado valor positivo (+1.00):

Si nos fijamos en el gráfico de esta situación  y en su complementario, podemos comprobar que la proporción que queda por debajo de +1.00 es el total (1 o el 100 %) menos la proporción que queda por encima de ese valor (que está en las tablas):

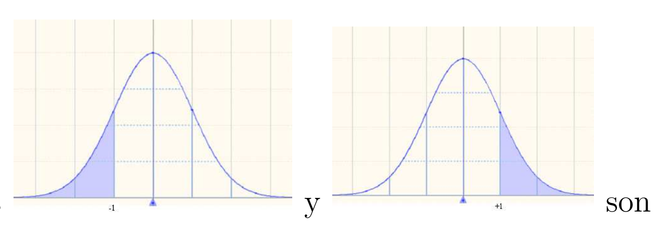
Por lo tanto: *La proporción por debajo de +1.00 es igual a 1-0.1587=0.8413*

b) Proporción de observaciones por encima de un determinado valor negativo (-1.00):

Como la curva es simétrica, las áreas  y son iguales, es decir, que la proporción por encima de (-1.00) es la misma que la que queda por debajo de (+1.00).

Por lo tanto: *La proporción por encima de -1.00 es igual a 1-0.1587=0.8413*

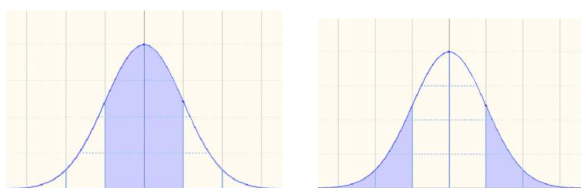
c) Proporción de observaciones por debajo de un determinado valor negativo (-1.00):

Como la curva es simétrica, las áreas  y son iguales, por lo que:

La proporción por debajo de -1.00 es igual a 0.1587.

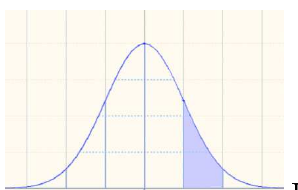
Ahora vamos a considerar la proporción de casos que se encuentran en un intervalo.

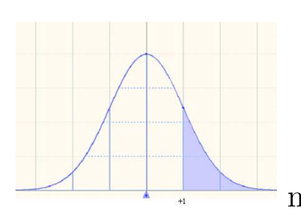
d) Proporción de datos entre dos valores simétricos respecto a la media: (-1.00 y +1.00):

Las áreas,  y son complementarias, y como la curva es simétrica, las dos ramas son iguales, por lo que la proporción de datos fuera del intervalo es el doble de la proporción de datos por encima de +1.00. Entonces:

La proporción de datos entre -1.00 y +1.00 es 1-2×0.1587 = 1-0.3174 = 0.6826.

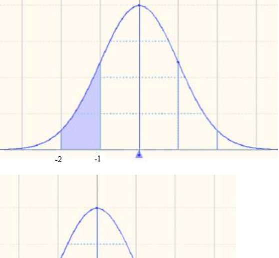
e) Proporción de datos entre dos valores positivos: (+1.00 y +2.00):

 Esta proporción la podemos expresar como la proporción de datos por encima de +1.00 menos la proporción de datos por encima de +2.00, es decir:

 menos . Entonces:

La proporción de datos entre +1.00 y +2.00 es 0.1587-0.0228=0.1359.

f) Proporción de datos entre dos valores negativos: (-2.00 y -1.00):

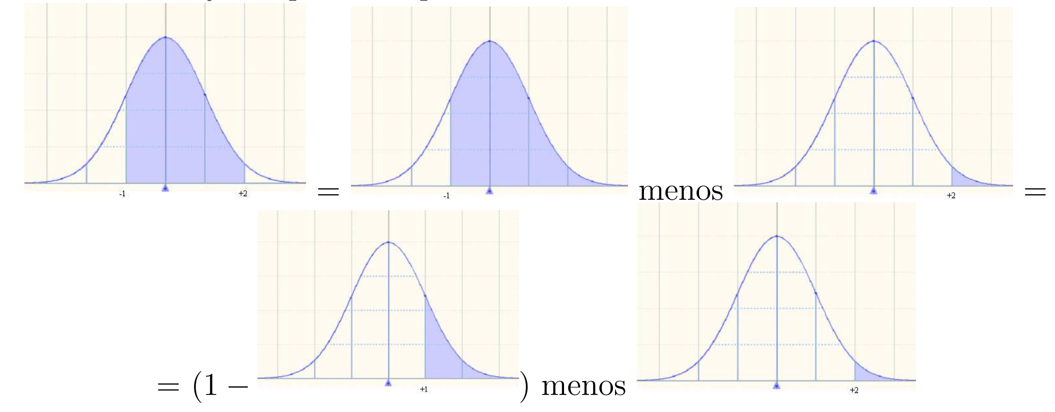


Como la curva es simétrica, la proporción de datos entre -2.00 y -1.00 es exactamente igual a la proporción de datos entre +1.00 y +2.00 .

La proporción de datos entre -2.00 y -1.00, es 0.1587-0.0228=0.1359.

g) Proporción de datos entre dos valores de distinto signo: (-1.00 y +2.00):

Usando los argumentos anteriores, esta proporción es la diferencia entre los que están por encima de -1.00 y los que están por encima de +2.00:



= menos =

= (1 −) menos

La proporción de datos entre -1.00 y +2.00, es (1-0.1587) - 0.0228=0.8413-0.0228=0.8185.

Recuerda que estamos usando todo el tiempo dos propiedades básicas de la curva Normal:

- Es simétrica respecto a la media.
- El área total bajo la curva vale 1.

4.3.3. Obtención de valores críticos

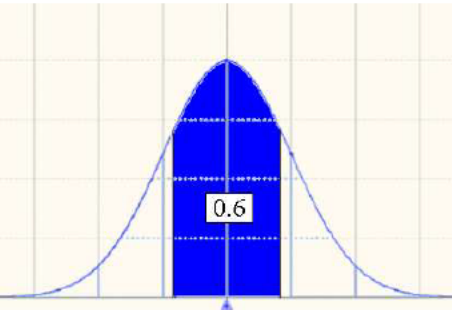
Del mismo modo que nos preguntamos por la proporción de observaciones que se encuentran en un determinado intervalo de valores tipificados, nos podríamos hacer la pregunta inversa: **¿cuál es el valor tipificado a partir del cual se encuentra una determinada proporción de observaciones?**

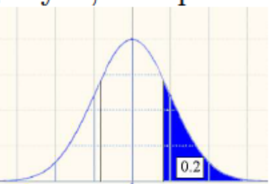
Podemos responder a esta cuestión utilizando la tabla de forma parecida al caso anterior.

Recordemos que las calificaciones del primer trabajo se distribuían según una Normal de media 6.5 y desviación típica 1.7, $N(6.5, 1.7)$.

Si queremos determinar entre qué puntuaciones se encuentra el 60 % central de las calificaciones, haremos lo siguiente:

Como la distribución de los valores observados y la de los valores tipificados son equivalentes, usaremos la tabla de la $N(0, 1)$ para obtener los valores que determinan ese intervalo y después desharemos el cambio.



Como la curva es simétrica, lo que necesitamos es el valor que deja un 30 % de las observaciones entre el origen y él, o lo que es lo mismo, el valor que deja a su derecha un 20 % de las observaciones 

, y su simétrico.

Buscamos en la tabla dicha proporción

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

El valor **más cercano** corresponde al valor típico: **0.84.**

(Si el valor buscado queda justo en medio de dos valores típicos, tomamos la media de ambos)

Esto significa que el 60 % central de las puntuaciones típicas se encuentran entre -0.84 y +0.84.

A nosotros nos interesan las calificaciones del trabajo y no los valores típicos, por lo que deberemos deshacer el cambio:

$$\text{Como } Z = \frac{X - \bar{x}}{s'}, \text{ entonces } X = s'Z + \bar{x}$$

En nuestro caso, el límite inferior será:

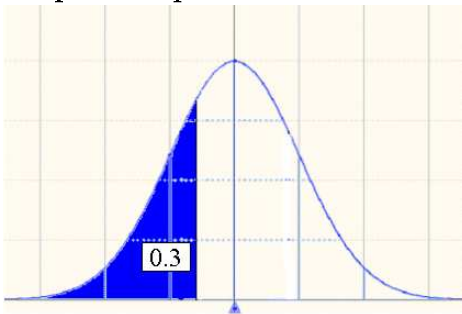
$$x_{inf} = s'z_{inf} + \bar{x} = 1.7 \times (-0.84) + 6.5 = 5.072$$

y el límite superior será:

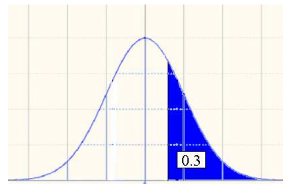
$$x_{sup} = s'z_{sup} + \bar{x} = 1.7 \times (0.84) + 6.5 = 7.928$$

Por lo tanto, el 60 % central de las calificaciones se encuentran entre 5.07 y 7.93.

Puntuación que no es superada por el 30 % de los estudiantes.



Estamos diciendo que el 30 % de los estudiantes tienen una nota inferior a esa.

Sabemos que buscamos un valor típico negativo pero, aprovechando la simetría, podemos buscar: 

.

A una proporción por encima de 0.3, le corresponde un valor típico de 0.52, entonces, el valor típico que buscamos es $z =$**-0.52** y la puntuación será:

$$x = s'z + \bar{x} = 1.7 \times (-0.52) + 6.5 = 5.616$$

Es decir, el 30 % de los estudiantes han obtenido una calificación inferior a 5.616 en el primer trabajo.

4.4. La distribución t de Student

Un modelo alternativo que se usará mucho en Inferencia Estadística es la llamada **distribución t de Student**.

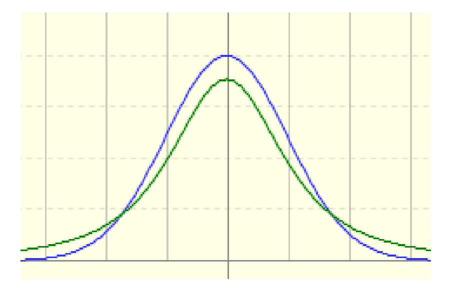
Esta curva es muy parecida a la curva Normal, $N(0,1)$, pero depende de un parámetro llamado «grados de libertad». Tiene, como la Normal, forma de campana, su media es cero y es simétrica, pero su varianza es mayor que uno.

Tiene la particularidad de que cuanto mayor es el parámetro grados de libertad, más se acerca la varianza a 1 y por lo tanto más se parece esta distribución a la distribución $N(0,1)$.

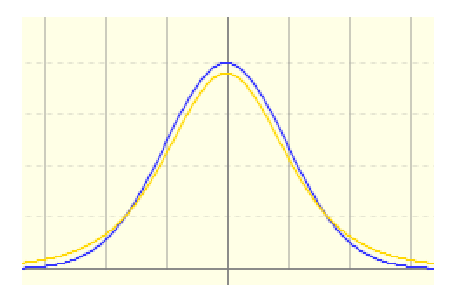
De hecho, cuando el número de grados de libertad es mayor que 30, la diferencia entre la t de Student y la $N(0,1)$ se puede considerar despreciable. Gráficamente:

http://www.matematicasvisuales.com/html/probabilidad/varaleat/tstudent.html

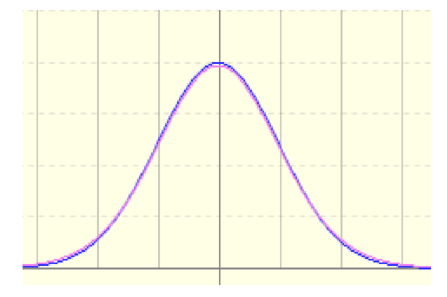
Comparación entre la gráfica de la $N(0,1)$ (azul) y la t de Student con 2 grados de libertad, t_2 (verde).



Comparación entre la gráfica de la $N(0,1)$ (azul) y la t de Student con 5 grados de libertad, t_5 (amarillo).

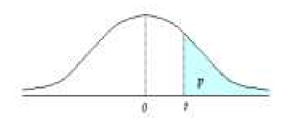


Comparación entre la gráfica de la $N(0,1)$ (azul) y la t de Student con 20 grados de libertad, t_{20} (rosa).



Las proporciones bajo la t de Student, se calculan como sigue:

Dada una proporción p y los grados de libertad n, la tabla nos proporciona el valor típico correspondiente:



$P\{T > t\} = p =$ Área sombreada

	p											
n	**0.005**	**0.01**	**0.025**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**
1	63.6567	31.8205	12.7062	6.3138	3.0777	1.9626	1.3764	1.0000	0.7265	0.5095	0.3249	0.1584
2	9.9248	6.9646	4.3027	2.9200	1.8856	1.3862	1.0607	0.8165	0.6172	0.4447	0.2887	0.1421
3	5.8409	4.5407	3.1824	2.3534	1.6377	1.2498	0.9785	0.7649	0.5844	0.4242	0.2767	0.1366
4	4.6041	3.7469	2.7764	2.1318	1.5332	1.1896	0.9410	0.7407	0.5686	0.4142	0.2707	0.1338
5	4.0321	3.3649	2.5706	2.0150	1.4759	1.1558	0.9195	0.7267	0.5594	0.4082	0.2672	0.1322
6	3.7074	3.1427	2.4469	1.9432	1.4398	1.1342	0.9057	0.7176	0.5534	0.4043	0.2648	0.1311
7	3.4995	2.9980	2.3646	1.8946	1.4149	1.1192	0.8960	0.7111	0.5491	0.4015	0.2632	0.1303
8	3.3554	2.8965	2.3060	1.8595	1.3968	1.1081	0.8889	0.7064	0.5459	0.3995	0.2619	0.1297
9	3.2498	2.8214	2.2622	1.8331	1.3830	1.0997	0.8834	0.7027	0.5435	0.3979	0.2610	0.1293
10	3.1693	2.7638	2.2281	1.8125	1.3722	1.0931	0.8791	0.6998	0.5415	0.3966	0.2602	0.1289
11	3.1058	2.7181	2.2010	1.7959	1.3634	1.0877	0.8755	0.6974	0.5399	0.3956	0.2596	0.1286
12	3.0545	2.6810	2.1788	1.7823	1.3562	1.0832	0.8726	0.6955	0.5386	0.3947	0.2590	0.1283
13	3.0123	2.6503	2.1604	1.7709	1.3502	1.0795	0.8702	0.6938	0.5375	0.3940	0.2586	0.1281
14	2.9768	2.6245	2.1448	1.7613	1.3450	1.0763	0.8681	0.6924	0.5366	0.3933	0.2582	0.1280
15	2.9467	2.6025	2.1314	1.7531	1.3406	1.0735	0.8662	0.6912	0.5357	0.3928	0.2579	0.1278
16	2.9208	2.5835	2.1199	1.7459	1.3368	1.0711	0.8647	0.6901	0.5350	0.3923	0.2576	0.1277
17	2.8982	2.5669	2.1098	1.7396	1.3334	1.0690	0.8633	0.6892	0.5344	0.3919	0.2573	0.1276
18	2.8784	2.5524	2.1009	1.7341	1.3304	1.0672	0.8620	0.6884	0.5338	0.3915	0.2571	0.1274
19	2.8609	2.5395	2.0930	1.7291	1.3277	1.0655	0.8610	0.6876	0.5333	0.3912	0.2569	0.1274
20	2.8453	2.5280	2.0860	1.7247	1.3253	1.0640	0.8600	0.6870	0.5329	0.3909	0.2567	0.1273
21	2.8314	2.5176	2.0796	1.7207	1.3232	1.0627	0.8591	0.6864	0.5325	0.3906	0.2566	0.1272
22	2.8188	2.5083	2.0739	1.7171	1.3212	1.0614	0.8583	0.6858	0.5321	0.3904	0.2564	0.1271
23	2.8073	2.4999	2.0687	1.7139	1.3195	1.0603	0.8575	0.6853	0.5317	0.3902	0.2563	0.1271
24	2.7969	2.4922	2.0639	1.7109	1.3178	1.0593	0.8569	0.6848	0.5314	0.3900	0.2562	0.1270
25	2.7874	2.4851	2.0595	1.7081	1.3163	1.0584	0.8562	0.6844	0.5312	0.3898	0.2561	0.1269
26	2.7787	2.4786	2.0555	1.7056	1.3150	1.0575	0.8557	0.6840	0.5309	0.3896	0.2560	0.1269
27	2.7707	2.4727	2.0518	1.7033	1.3137	1.0567	0.8551	0.6837	0.5306	0.3894	0.2559	0.1268
28	2.7633	2.4671	2.0484	1.7011	1.3125	1.0560	0.8546	0.6834	0.5304	0.3893	0.2558	0.1268
29	2.7564	2.4620	2.0452	1.6991	1.3114	1.0553	0.8542	0.6830	0.5302	0.3892	0.2557	0.1268
30	2.7500	2.4573	2.0423	1.6973	1.3104	1.0547	0.8538	0.6828	0.5300	0.3890	0.2556	0.1267
31	2.7440	2.4528	2.0395	1.6955	1.3095	1.0541	0.8534	0.6825	0.5298	0.3889	0.2555	0.1267
32	2.7385	2.4487	2.0369	1.6939	1.3086	1.0535	0.8530	0.6822	0.5297	0.3888	0.2555	0.1267
33	2.7333	2.4448	2.0345	1.6924	1.3077	1.0530	0.8526	0.6820	0.5295	0.3887	0.2554	0.1266
34	2.7284	2.4411	2.0322	1.6909	1.3070	1.0525	0.8523	0.6818	0.5294	0.3886	0.2553	0.1266
35	2.7238	2.4377	2.0301	1.6896	1.3062	1.0520	0.8520	0.6816	0.5292	0.3885	0.2553	0.1266
36	2.7195	2.4345	2.0281	1.6883	1.3055	1.0516	0.8517	0.6814	0.5291	0.3884	0.2552	0.1266
37	2.7154	2.4314	2.0262	1.6871	1.3049	1.0512	0.8514	0.6812	0.5289	0.3883	0.2552	0.1265
38	2.7116	2.4286	2.0244	1.6860	1.3042	1.0508	0.8512	0.6810	0.5288	0.3882	0.2551	0.1265
39	2.7079	2.4258	2.0227	1.6849	1.3036	1.0504	0.8509	0.6808	0.5287	0.3882	0.2551	0.1265
40	2.7045	2.4233	2.0211	1.6839	1.3031	1.0500	0.8507	0.6807	0.5286	0.3881	0.2550	0.1265
45	2.6896	2.4121	2.0141	1.6794	1.3006	1.0485	0.8497	0.6800	0.5281	0.3878	0.2549	0.1264
50	2.6778	2.4033	2.0086	1.6759	1.2987	1.0473	0.8489	0.6794	0.5278	0.3875	0.2547	0.1263
55	2.6682	2.3961	2.0040	1.6730	1.2971	1.0463	0.8482	0.6790	0.5275	0.3873	0.2546	0.1262
60	2.6603	2.3901	2.0003	1.6706	1.2958	1.0455	0.8477	0.6786	0.5272	0.3872	0.2545	0.1262
65	2.6536	2.3851	1.9971	1.6686	1.2947	1.0448	0.8472	0.6783	0.5270	0.3870	0.2544	0.1262
70	2.6479	2.3808	1.9944	1.6669	1.2938	1.0442	0.8468	0.6780	0.5268	0.3869	0.2543	0.1261
75	2.6430	2.3771	1.9921	1.6654	1.2929	1.0436	0.8464	0.6778	0.5266	0.3868	0.2542	0.1261
80	2.6387	2.3739	1.9901	1.6641	1.2922	1.0432	0.8461	0.6776	0.5265	0.3867	0.2542	0.1261
85	2.6349	2.3710	1.9883	1.6630	1.2916	1.0428	0.8459	0.6774	0.5264	0.3866	0.2541	0.1260
90	2.6316	2.3685	1.9867	1.6620	1.2910	1.0424	0.8456	0.6772	0.5263	0.3866	0.2541	0.1260
95	2.6286	2.3662	1.9853	1.6611	1.2905	1.0421	0.8454	0.6771	0.5262	0.3865	0.2541	0.1260
100	2.6259	2.3642	1.9840	1.6602	1.2901	1.0418	0.8452	0.6770	0.5261	0.3864	0.2540	0.1260
125	2.6157	2.3565	1.9791	1.6571	1.2884	1.0408	0.8445	0.6765	0.5257	0.3862	0.2539	0.1259
150	2.6090	2.3515	1.9759	1.6551	1.2872	1.0400	0.8440	0.6761	0.5255	0.3861	0.2538	0.1259
200	2.6006	2.3451	1.9719	1.6525	1.2858	1.0391	0.8434	0.6757	0.5252	0.3859	0.2537	0.1258
300	2.5923	2.3388	1.9679	1.6499	1.2844	1.0382	0.8428	0.6753	0.5250	0.3857	0.2536	0.1258
∞	2.5758	2.3263	1.9600	1.6449	1.2816	1.0364	0.8416	0.6745	0.5244	0.3853	0.2533	0.1257

Si observamos la tabla, veremos que con ella podemos hacer aún menos aproximaciones que en el caso de la Normal (hay muy pocos valores de p).

Veamos algunos ejemplos:

1. Determina la proporción de observaciones que están por encima del valor 1 para una t de Student con 16 grados de libertad.

   Buscamos en la fila correspondiente a n=16 el valor más próximo a 1 (1.0711), que nos da una p=0.15. Entonces: la proporción buscada es **0.15** , el 15 % de las observaciones.

2. Determina la proporción de observaciones que están por debajo del valor 2.53 para una t de Student con 20 grados de libertad.

   Buscamos en la fila correspondiente a n=20 el valor más próximo a 2.53 (2.5280), que nos da una p=0.01. Esto significa que 0.01 es la proporción de observaciones por encima de dicho valor. Entonces: la proporción buscada es **1-0.01=0.99**, es decir el 99 % de las observaciones.

3. Determina la proporción de observaciones que están por encima del valor -0.7 para una t de Student con 7 grados de libertad.

   Para los valores negativos aprovecharemos la simetría de la gráfica: el área por encima de -0.7 es igual al área por debajo de 0.7, entonces:

   Buscamos en la fila correspondiente a n=7 el valor más próximo a 0.7 (0.7111), que nos da una p=0.25. Esto significa que 0.25 es la proporción de observaciones por encima de dicho valor. Entonces: la proporción buscada es **1-0.25=0.75**, es decir el 75 % de las observaciones.

4. Determina qué valor de una t de Student con 50 grados de libertad deja a su derecha un área de 0.25.

   Buscamos en la fila de n=50 la intersección con la columna p=0.25 y obtenemos el valor buscado: **0.6794**.

5. Determina qué valor de una t de Student con 22 grados de libertad verifica que el área encerrada entre este valor y 0.2564 es exactamente 0.1.

   Si hacemos el dibujo (siempre ayuda mucho), podemos observar que el área por encima del valor buscado es igual a 0.1 más el área por encima de 0.2564.

   El área por encima de 0.2564 en una t de Student con 22 grados de libertad es 0.4, y por lo tanto, el área por encima del valor buscado es 0.4+0.1=0.5. Esto significa que el valor que estamos buscando es **cero**.

6. Determina qué valor de una t de Student con 40 grados de libertad verifica que el área encerrada entre -1.05 y este valor es exactamente 0.7.

   Volvemos al dibujo. Podemos observar que el área por **debajo** del valor buscado es igual a 0.7 más el área por debajo de -1.05.

Aprovechando la simetría, sabemos que el área por debajo de -1.05 es igual al área por encima de 1.05. Buscamos en la tabla dicha área para una t de Student con 40 grados de libertad y obtenemos p=0.15. Entonces:

El área por debajo del valor buscado es igual a 0.7+0.15=0.85, lo que significa que el área por encima es 0.15. Por lo tanto el valor buscado es: **1.05**.

Notas:

1. Siempre que hagamos cálculos con las tablas (tanto de la Normal como de la t de Student), es muy recomendable hacer el dibujo correspondiente para entender lo que calculamos y no equivocarnos.

2. Los programas estadísticos sí que nos permiten obtener las probabilidades o los valores críticos en cualquier situación.

Tema 5

Probabilidad y variables aleatorias

Vamos a intentar abordar ahora una situación no determinista.

En la mayoría de las situaciones con las que trabajaremos, vamos a tener que tomar decisiones y sacar conclusiones aceptando un cierto riesgo, un cierto nivel de incertidumbre que viene dado por el hecho de que en nuestro estudio no podemos predecir exactamente el resultado de un experimento o no podemos realizarlo tantas veces como sería deseable.

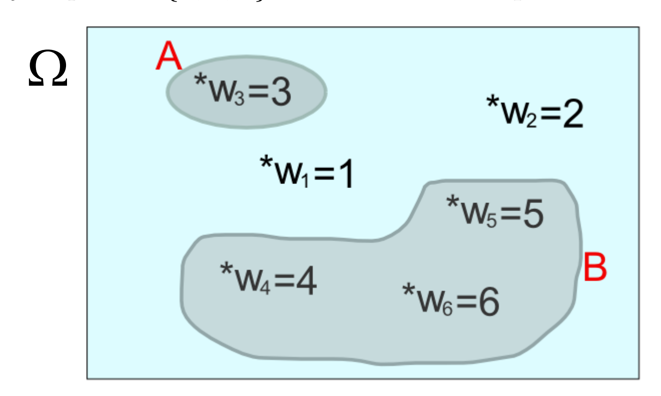
Si lanzamos una moneda al aire (no trucada), no sabemos qué va a ocurrir. Sin embargo, nos interesa poder describir cuál es el comportamiento de los resultados del experimento «lanzar una moneda».

Vamos a introducir algunas definiciones que nos permitan abordar estas situaciones para poder describirlas.

A los experimentos de este tipo, en los cuales no se puede predecir cuál va a ser el resultado, se les denomina **experimentos aleatorios**, a cada uno de los posibles resultados del mismo se le denomina **suceso elemental** y al conjunto de todos los posibles resultados del experimento, se le denomina **espacio muestral** y se le suele denotar por **E** o por Ω.

- Cada subconjunto del espacio muestral es un suceso, y puede ser elemental o compuesto.

  Ejemplo: si lanzamos un dado, A=*sacar un 3*=$\{3\}$ es un suceso elemental y B=*sacar un número mayor que 3*=$\{4, 5, 6\}$ es un suceso compuesto.



- Al suceso que ocurre siempre se le llama **suceso seguro** y coincide con el espacio muestral.

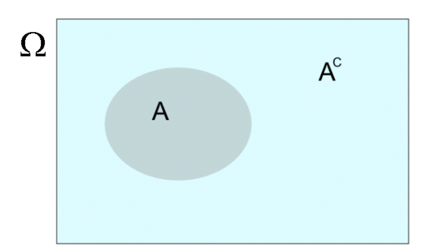
  Ejemplo: $A =$*sacar un número del 1 al 6 al lanzar un dado*$=\Omega$.

- Al suceso que no puede ocurrir nunca se le llama **suceso imposible** y se denota por el vacío ($\emptyset$).

  Ejemplo: $A =$*sacar un 8 al lanzar un dado*$=\emptyset$.

- Llamaremos **suceso contrario o complementario** de un suceso A, a lo que ocurre cuando no ocurre A.

  Ejemplo: si A es el suceso *sacar un 3* al lanzar un dado: A=$\{3\}$, entonces A^c=*no sacar un 3*, es decir: A^c=$\{1,2,4,5,6\}$

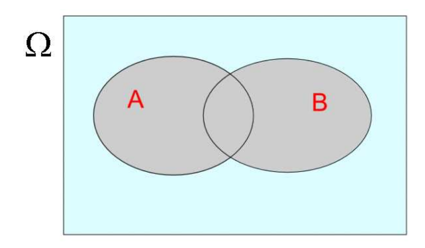
  Es decir: $A^c = \Omega - A$

5.1. Operaciones con sucesos

Entre los sucesos se pueden establecer las siguientes operaciones:

Unión de sucesos

Dados dos sucesos A y B, llamaremos suceso unión de A y B, al suceso formado por todos los sucesos elementales de A y de B:



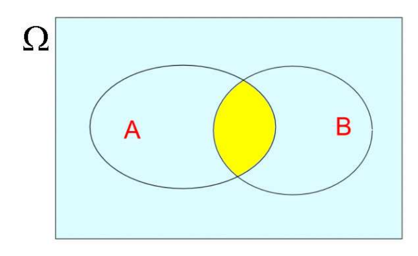
Es el suceso que ocurre cuando ocurre A **o** B **o** los dos y se denota como $A \bigcup B$.

Ejemplo: si al lanzar un dado, $A = \{3, 4\}$ y $B =$ *sacar un número par* $= \{2, 4, 6\}$,

entonces $A \bigcup B = \{2, 3, 4, 6\}$

Intersección de sucesos

Dados dos sucesos A y B, llamaremos suceso intersección de A y B, al suceso formado por todos los sucesos elementales comunes a A y a B:



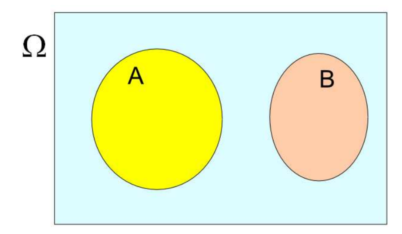
Es el suceso que ocurre cuando ocurren A **y** B a la vez y se denota como $A \bigcap B$.

Ejemplo: si al lanzar un dado, $A = \{3, 4\}$ y $B = \textit{sacar un número par} = \{2, 4, 6\}$,

entonces $A \bigcap B = \{4\}$.

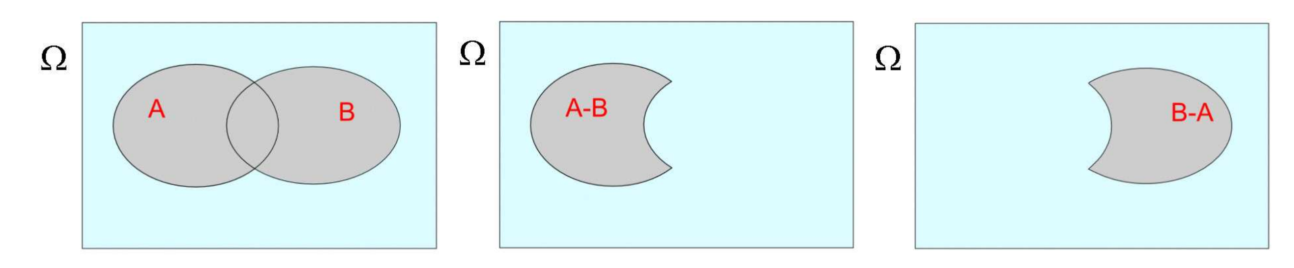
Los sucesos que no pueden ocurrir a la vez se llaman **sucesos incompatibles** y su intersección es el suceso imposible.

Ejemplo: si al lanzar un dado, $A = \{3\}$ y $B = \textit{sacar un número par} = \{2, 4, 6\}$, entonces $A \bigcap B = \emptyset$.



Diferencia de sucesos

Dados dos sucesos A y B, llamaremos diferencia $A - B$, al suceso formado por todos los sucesos elementales de A que no están en B.



Es el suceso que ocurre cuando ocurre A pero no ocurre B: $A - B = A \bigcap B^c$

Ejemplo: si al lanzar un dado, A={3,4} y B=*sacar un número par*={2,4,6}, entonces

$$A - B = \{3\} \text{ y } B - A = \{2, 6\}$$

Es interesante observar que: $A - B = A - (A \bigcap B)$

5.2. Probabilidad

Al estudiar los experimentos aleatorios, aunque no sabemos cuál va a ser el resultado, sí que el sentido común, o la intuición, nos indica que hay unos resultados que tienen más posibilidades de ocurrir que otros. Vamos, entonces, a intentar plasmar esta idea intuitiva. Es decir, vamos a determinar la **probabilidad de que ocurra un determinado suceso**.

Una forma de obtener la probabilidad «teórica», no basada en los resultados del experimento, es la llamada **probabilidad clásica o Regla de Laplace**.

Si todos los resultados de un experimento tienen la misma posibilidad de ocurrir, entonces, **la probabilidad de ocurrencia de un suceso A será**:

$$P(A) = \frac{\text{número de casos favorables a } A}{\text{número de casos posibles}}$$

Ejemplo: Consideremos el experimento *lanzar un dado.*

Nuestro espacio muestral es Ω=$\{1, 2, 3, 4, 5, 6\}$ y, si el dado no está trucado, suponemos que todos los sucesos tienen la misma posibilidad de ocurrir. Entonces podemos calcular:

$$P(5) = \frac{\text{número de casos favorables a 5}}{\text{número de casos posibles}} = \frac{1}{6} = 0.1\hat{6}$$

$$P(par) = \frac{\text{número de casos favorables a } par}{\text{número de casos posibles}} = \frac{3}{6} = 0.5$$

Otra forma de obtener una probabilidad es la **probabilidad «empírica» o frecuentista**.

Una vez comprobados algunos resultados experimentales, se define la **probabilidad de ocurrencia de un suceso** A como:

$$P(A) = \frac{\text{número de veces que ha ocurrido } A}{\text{número de repeticiones}}$$

(es la frecuencia relativa de A).

Cuando el número de repeticiones es bajo, esta probabilidad «empírica» puede ser bastante incorrecta, mientras que, a medida que aumentamos el número de repeticiones, el valor de la probabilidad se estabiliza y se va aproximando, cada vez más a la probabilidad «teórica».

Ejemplo: si lanzamos un dado y contamos las veces que ha salido un 2:

Nº de lanzamientos	50	100	150	200	300	400	500
Nº de 2	4	13	24	34	51	64	80
P(2)= frecuencia relativa	0.08	0.13	0.16	0.17	0.17	0.16	0.16

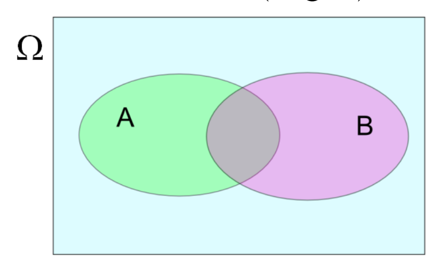
Estas ideas se pueden formalizar dando la siguiente **definición axiomática de la probabilidad**:

Una **distribución de probabilidad** es una función que asigna a cada suceso posible un número en el intervalo [0,1], con las siguientes propiedades:

1. Para todo suceso $A : P(A) \geq 0$
2. La probabilidad del suceso seguro es 1: $P(\Omega) = 1$
3. Si A y B son dos sucesos incompatibles ($A \bigcap B = \emptyset$), entonces:
   $P(A \bigcup B) = P(A) + P(B)$

De esta definición se deducen las siguientes consecuencias:

1. Si A^c es el suceso contrario de A, entonces: $P(A^c) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. Dados dos sucesos A y B cualesquiera: $P(A \bigcup B) = P(A) + P(B) - P(A \bigcap B)$



4. Podemos generalizar el axioma 3: Si $A = A_1 \bigcup A_2 \bigcup ... \bigcup A_n$, siendo estos sucesos incompatibles 2 a 2, entonces:
   $P(A) = P(A_1 \bigcup A_2 \bigcup ... \bigcup A_n) = P(A_1) + P(A_2) + ... + P(A_n)$

5.3. Probabilidades condicionadas

Nos podemos plantear cuál será la probabilidad de cierto suceso B, sabiendo que ha sucedido otro suceso A. Por ejemplo, se lanza un dado y nos dicen que el resultado es impar ¿cuál es la probabilidad de que sea un 3?

Esta probabilidad se conoce como **probabilidad condicionada**, y se calcula de la siguiente manera:

La probabilidad del segundo suceso, B, dado que conocemos que ha ocurrido el primer suceso, A, o bien, la probabilidad del suceso B condicionado a que ha ocurrido el suceso A es:

$$P(B|A) = \frac{P(A\bigcap B)}{P(A)}, \text{siendo } P(A) > 0$$

Ejemplo:

$$P(3|impar) = \frac{P(3 \text{ e } impar)}{P(impar)} = \frac{P(3)}{P(impar)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

De aquí se deduce la fórmula de la **probabilidad compuesta**: La probabilidad de que ocurran dos sucesos A y B simultáneamente es la probabilidad de que ocurra uno por la probabilidad de que ocurra el otro dado que ha ocurrido el primero:

$$P(A\bigcap B) = P(A)P(B|A) = P(B)P(A|B)$$

Diremos que dos sucesos son **independientes** cuando la ocurrencia del primero no cambia la probabilidad de que ocurra el segundo.

$$P(B \mid A) = P(B) \text{ , o bien } P(A \mid B) = P(A)$$

Como consecuencia, si dos sucesos son independientes: $P(A\bigcap B) = P(A)P(B)$

5.4. Variables aleatorias

Cuando realizamos un experimento, tenemos un espacio muestral (con todos los resultados, sucesos elementales, posibles).

Una **variable aleatoria**, es una función que asocia a cada suceso elemental un número perfectamente definido.

$$\zeta : \Omega \longrightarrow \mathbb{R}$$

Por ejemplo, en el caso de lanzar 2 monedas, podemos estudiar el número de caras y entonces asociar números a los resultados:

$$(c, c) \to 2; (c, x) \to 1; (x, c) \to 1; (x, x) \to 0$$

Llamaremos **función de distribución de una variable aleatoria** a una función F, $F : \mathbb{R} \longrightarrow [0, 1]$, que asocia a cada valor x la probabilidad de que la variable aleatoria tome un valor **menor o igual** que x: $F(x) = P(\zeta \leq x)$.

$F(x)$, es la «probabilidad **acumulada**» hasta x.

Ejemplo: en el caso anterior, $F(0) = 1/4$; $F(1) = 3/4$; $F(2) = 1$

En general: $P(a < \zeta \leq b) = F(b) - F(a)$

Para seguir con el análisis, debemos distinguir dos tipos de variables aleatorias: las discretas y las continuas.

Llamaremos **variable aleatoria discreta**, a una variable aleatoria cuyo soporte (conjunto de valores posibles) es un *conjunto discreto* (finito o numerable).

Ejemplo: la variable aleatoria anterior es discreta. Su soporte es el conjunto $\{0, 1, 2\}$, que es un conjunto finito.

Llamaremos **variable aleatoria continua**, a una variable aleatoria cuyo soporte (conjunto de valores posibles) NO es un conjunto discreto (intuitivamente, este conjunto será entonces un intervalo de números reales).

Ejemplo: la variable aleatoria que asigna a cada persona extraída de una población su peso, es una variable aleatoria continua ya que podemos considerar como posibles todos los valores del intervalo $(0, 300)$.

En el caso de las variables aleatorias discretas, vamos a construir una función asociando a cada uno de los valores de la variable aleatoria, su probabilidad:

Si tenemos una variable aleatoria discreta ζ, que toma los valores $x_1, \ldots, x_n$, entonces:

$f(x_i) = P(\zeta = x_i) = p_i$.

Además se cumple que: $f(x_1) + f(x_2) + \cdots + f(x_n) = 1$

A esta función f que acabamos de construir, se le llama **función de probabilidad o función de cuantía** de una variable aleatoria de tipo discreto.

Ejemplo: en el caso anterior, $f(0) = 1/4$; $f(1) = 2/4 = 1/2$; $f(2) = 1/4$

Ejemplo: Si lanzamos un dado al aire, la variable aleatoria asociada a este experimento tomará los valores: $1, 2, 3, 4, 5, 6$ y la probabilidad de cada uno de estos resultados es $1/6$.

La función de probabilidad f, es tal que:

$f(1) = P(\zeta = 1) = 1/6$	$f(2) = P(\zeta = 2) = 1/6$	$f(3) = P(\zeta = 3) = 1/6$
$f(4) = P(\zeta = 4) = 1/6$	$f(5) = P(\zeta = 5) = 1/6$	$f(6) = P(\zeta = 6) = 1/6$

Y se cumple que: $f(1) + f(2) + \cdots + f(6) = 1$

Para cualquier otro valor, la función de probabilidad vale cero.

$$f(2.5) = P(\zeta = 2.5) = 0$$

Si calculamos la función de distribución F para los valores anteriores, obtenemos:

$F(1) = P(\zeta \leq 1) = 1/6$	$F(2) = P(\zeta \leq 2) = 2/6$	$F(3) = P(\zeta \leq 3) = 3/6$
$F(4) = P(\zeta \leq 4) = 4/6$	$F(5) = P(\zeta \leq 5) = 5/6$	$F(6) = P(\zeta \leq 6) = 6/6 = 1$

En el resto de los puntos, se calcula de forma análoga:

$$F(2.5) = P(\zeta \leq 2.5) = P(\zeta = 1) + P(\zeta = 2) = 2/6 = 1/3$$

Entonces, para conocer cuál es la probabilidad de que al lanzar un dado obtengamos una puntuación mayor que 2 y menor o igual que 5, podríamos hacerlo de dos formas:

Mediante la función de probabilidad:

$$P(2 < \zeta \leq 5) = P(\zeta = 3) + P(\zeta = 4) + P(\zeta = 5) = 3 \times 1/6 = 1/2$$

O usando la función de distribución:

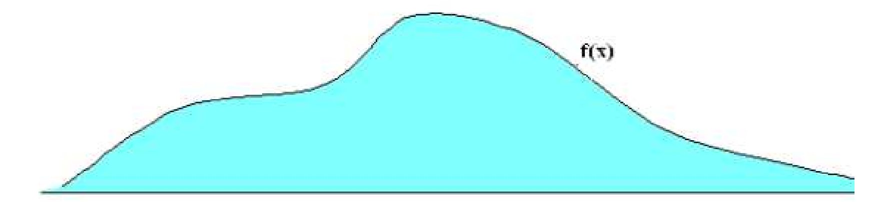
$$P(2 < \zeta \leq 5) = F(5) - F(2) = 5/6 - 2/6 = 3/6 = 1/2$$

Cuando observamos tiempos, longitudes, etc..., la variable aleatoria resultante es una **variable aleatoria continua**.

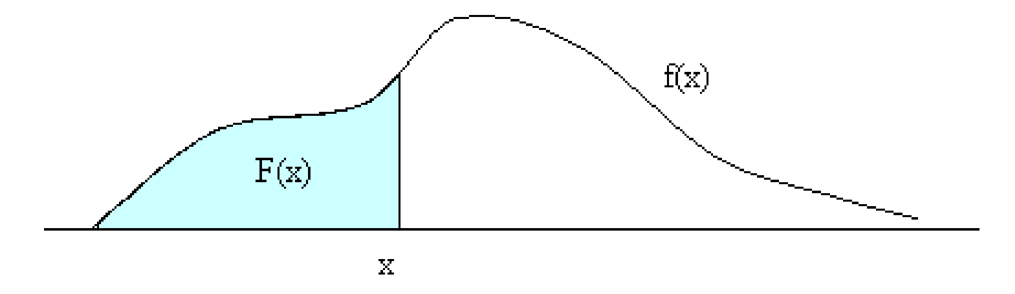
En este caso, en lugar de utilizar la función de probabilidad o de cuantía, se usa la llamada **función de densidad de probabilidad**, que es una función tal que el área comprendida bajo la curva, entre dos puntos, es precisamente la probabilidad entre esos dos puntos.

Para que una función pueda ser **la función de densidad de una variable aleatoria continua** tiene que cumplir :

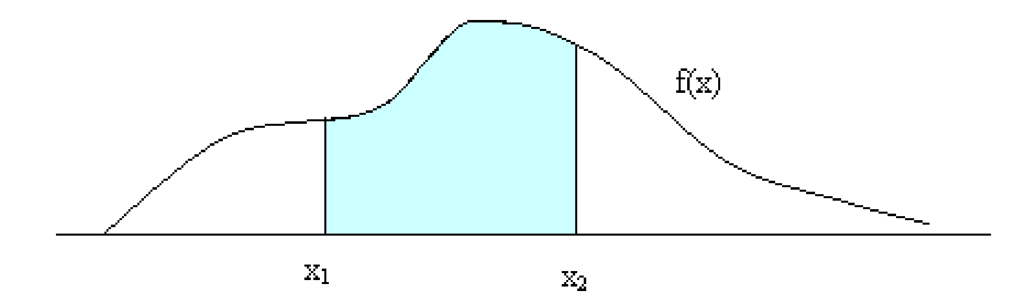
- $f(x) \geq 0$, para cualquier valor x de la variable.
- El área total encerrada entre el eje horizontal y la curva $f(x)$, vale 1.



Sabemos que $F(x) = P(\zeta \leq x)$. Gráficamente, $F(x)$ es el área encerrada bajo la curva $f(x)$, desde $-\infty$ hasta x:



Como consecuencia, $P(x_1 < \zeta \leq x_2) = F(x_2) - F(x_1)$



Nota: en el caso de las variables continuas, la probabilidad de que la variable aleatoria tome un valor concreto es cero ($P(\zeta = x) = 0$), y en consecuencia:

$$P(x_1 < \zeta \leq x_2) = P(x_1 < \zeta < x_2) = F(x_2) - F(x_1)$$

sin embargo, esto no es cierto en el caso de las variables aleatorias discretas (utilizamos aquí los datos del último ejemplo):

$$P(2 < \zeta \leq 5) = P(\zeta = 3) + P(\zeta = 4) + P(\zeta = 5) = 3 \times 1/6 = 1/2$$

mientras que:

$$P(2 < \zeta < 5) = P(\zeta = 3) + P(\zeta = 4) = 2 \times 1/6 = 1/3$$

5.5. Esperanza matemática

El concepto de esperanza matemática, como otros muchos de la teoría de la probabilidad, tiene su origen en los juegos de azar. Los jugadores deseaban conocer cuál era su esperanza de ganancias o pérdidas cuando participaban repetidamente en un juego. En este sentido, el valor esperado representa la cantidad de dinero promedio que el jugador espera ganar o perder después de un gran número de partidas.

Supongamos el siguiente juego:

El jugador lanza un dado, y: si sale un 1 el jugador gana 1 euro, si sale un 2 gana 4 euros, si sale un 3 gana 5 euros, si sale un 4 no gana ni pierde nada, si sale un 5 pierde 2 euros y si sale un 6 pierde 6 euros.

Vamos a calcular el valor esperado de las ganancias o pérdidas en el juego:

La variable aleatoria *ganancias en el juego* toma los valores {-6, -2, 0, 1, 4, 5}, y como los valores del dado son equiprobables, la probabilidad de cada uno de estos valores es 1/6.

Valor del dado	1	2	3	4	5	6
Probabilidad	1/6	1/6	1/6	1/6	1/6	1/6
Ganancia	1	4	5	0	-2	-6

Entonces, la ganancia esperada, después de un número grande de partidas, se obtiene sumando los productos de cada valor de la ganancia por la probabilidad de obtenerla. Es decir, en este juego:

Ganancia esperada=$1 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 0 \times \frac{1}{6} + (-2) \times \frac{1}{6} + (-6) \times \frac{1}{6} = 2 \times \frac{1}{6} = \frac{1}{3} = 0.\hat{3}$ euros

La ganancia esperada puede ser positiva, negativa o cero; en el primer caso diremos que el juego es favorable al jugador, en el segundo que es un juego desfavorable y si la ganancia esperada es cero diremos que es un juego justo.

Es importante destacar que el valor de la esperanza no tiene por qué ser un valor posible de la variable, lo que significa que una variable aleatoria puede que nunca tome el valor de su esperanza.

Este concepto también se utiliza en situaciones que nada tienen que ver con los juegos de azar, así podemos hablar de la *esperanza de vida de las mujeres* o del *tiempo esperado de permanencia en una consulta*. Estos valores esperados hay que interpretarlos como un promedio y no se pueden aplicar a un individuo en particular.

Si trasladamos este concepto a las variables aleatorias, podemos **interpretar la esperanza matemática de una variable aleatoria**, como su promedio o valor esperado, después de realizar un gran número de pruebas del experimento al que está asociada, de modo que tenemos la siguiente definición (la damos solo en el caso de una variable aleatoria discreta):

Si consideramos una variable aleatoria de tipo discreto, ζ, con función de probabilidad $p_i = P(\zeta = x_i)$, para todo $i = 1, \ldots n$, entonces definimos la **esperanza matemática, o valor esperado, o media** de ζ, como:

$$E[\zeta] = \mu = \sum_{i=1}^{n} x_i p_i$$

(Podemos entender la esperanza como un promedio de la variable)

Del mismo modo que hemos definido la esperanza, podemos definir la **varianza de una variable aleatoria** de tipo discreto:

Dada una variable aleatoria de tipo discreto, ζ, que tiene función de probabilidad $p_i = P(\zeta = x_i)$, para todo $i = 1, \ldots n$, entonces llamamos **varianza** de ζ, al valor:

$$\mathrm{Var}(\zeta) = \sigma^2 = E[(\zeta - \mu)^2]$$

Se puede comprobar que se cumple:

$$\mathrm{Var}(\zeta) = \sigma^2 = \sum_{i=1}^{n} x_i^2 p_i - \mu^2 = E[\zeta^2] - E[\zeta]^2$$

Llamaremos **desviación típica** de ζ a la raíz cuadrada positiva de la varianza.

$$\sigma = \mathrm{DT}(\zeta) = +\sqrt{\sigma^2}$$

5.6. La probabilidad y la curva Normal

Muchas de las funciones de densidad de variables aleatorias de tipo continuo tienen como representación gráfica la campana de Gauss, son las llamadas distribuciones normales. Una distribución Normal, está determinada cuando se conocen su media (μ) y su desviación típica (σ), y se denota por: $N(\mu, \sigma)$.

La **función de densidad $f(x)$, de una distribución Normal**, viene dada por la siguiente expresión (cuya gráfica es la campana de Gauss):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

donde:

$f(x)$ es la densidad de un determinado valor

x es un valor cualquiera de la variable

μ es la media de la distribución

σ es la desviación típica de la distribución

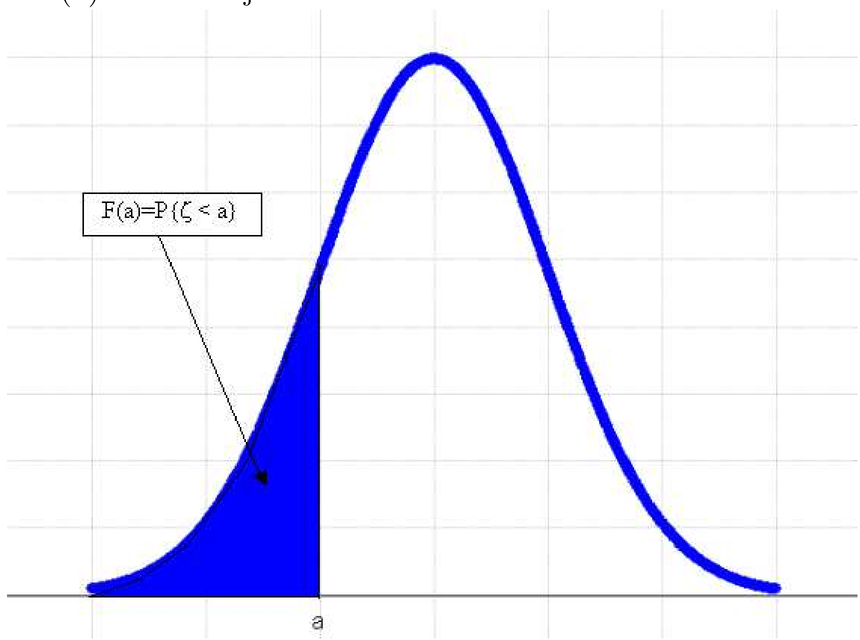
π es la constante: 3.14159...

e es la constante: 2.71828...

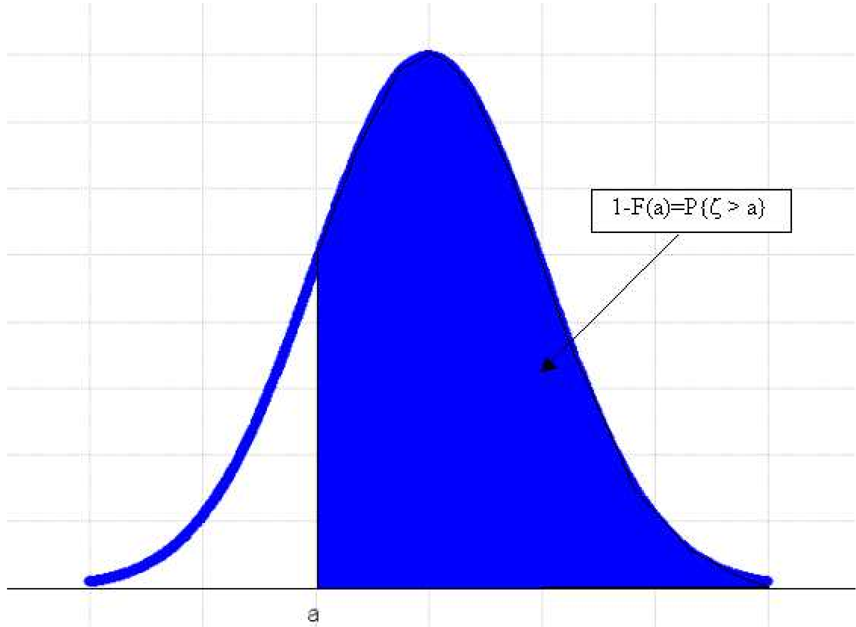
Recordemos que esta función está definida y es continua en $(-\infty, +\infty)$, es simétrica respecto a la media (μ), tiene un máximo en $x = \mu$ y el eje de abscisas es una asíntota horizontal (por mucho que se acerque en los extremos, la curva nunca llega a tocar el eje).

Para obtener la $P(\zeta \leq a)$, utilizaremos la función de distribución:

$P(\zeta \leq a) = F(a) =$ área bajo la curva hasta **a**:



Calcular la probabilidad mediante áreas no es fácil, pero ya conocemos las tablas de la distribución $N(0,1)$ que nos dan la proporción de observaciones por encima de un determinado valor **a**, lo que en **términos de variables aleatorias significa que nos dan la probabilidad de que la variable aleatoria tome un valor por encima de a**.



Entonces, **para calcular las probabilidades de una variable aleatoria que siga cualquier distribución** $N(\mu, \sigma)$, en primer lugar, tipificaremos (haciendo un cambio de variable para obtener otra variable aleatoria también Normal, pero con media 0 y desviación típica 1), y a continuación usaremos las tablas de la $N(0,1)$ que ya conocemos.

Tema 6

Introducción a la Inferencia Estadística

Como ya comentamos al principio de la asignatura, con la Estadística no solo queremos describir el comportamiento de una variable o característica de una población, sino que también la utilizaremos para tomar decisiones respecto a toda la población basándonos en los resultados obtenidos para una muestra.

Si en lugar de trabajar con toda la población estamos trabajando con una muestra, es muy importante que no confundamos las características de una y otra, así la media y la desviación típica muestrales las denotaremos por $\bar{x}$ y s', mientras que la media y la desviación típica poblacionales las denotaremos por μ y σ respectivamente.

6.1. Distribución de la media muestral

Vamos a comenzar planteando el tema de la estimación de la media (μ) de una variable para una población.

La población es tan grande que no podemos abordarla en su totalidad, y por lo tanto, solo podemos trabajar con los resultados obtenidos a partir de una muestra.

Supongamos que queremos determinar la altura media de los estudiantes de la Universidad de La Rioja. No vamos a poder medirlos a todos, así que elegimos, al azar, a 30 estudiantes, los medimos y calculamos la media de las alturas. Obtenemos como altura media: 177.32 cm, con una cuasidesviación típica (para estos 30 datos) de 12.98 cm.

¿Nos atreveríamos a decir que la altura media de los estudiantes de la Universidad de La Rioja es de 177.32 cm?

¿Qué ocurre si tomamos otras muestras, también de tamaño 30?

Muestra	1	2	3	4	5	6	7	8	9	10
Media	177.32	175.06	178.30	178.26	179.47	173.61	175.83	179.18	180.12	177.18
Cuasidesviación típica	12.98	11.96	11.65	13.46	9.75	13.03	13.32	13.14	12.12	11.01

Como podemos observar, para cada muestra obtenemos unos valores diferentes, pero estos valores, que podemos considerar como observaciones de la variable aleatoria *media muestral*, siguen también una distribución.

Se puede demostrar que la variable aleatoria *media muestral* (esa variable cuyos valores son las medias obtenidas de cada una de las posibles muestras de tamaño n), tiene la siguiente distribución:

Si tenemos una población en la que la variable considerada sigue una distribución Normal, $N(\mu, \sigma)$, y extraemos muestras de tamaño n, entonces, la variable aleatoria media muestral, sigue una distribución:

$$\bar{\zeta} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Notad que estamos diciendo que la variable aleatoria media muestral tiene como media (esperanza) la media poblacional (μ). **El valor esperado de la media muestral es la media poblacional**.

La desviación típica de la *media muestral* se conoce como error estándar o error típico de la media (standard error of the mean: SE).

El resultado anterior también es cierto (aproximadamente) cuando la distribución en la población no es Normal, siempre que el tamaño de las muestras sea suficientemente grande:

Teorema central del límite: Si se toman muestras de tamaño n (n>30) de una población con una distribución cualquiera, de media μ y desviación típica σ, entonces, la distribución de la variable aleatoria media muestral sigue, aproximadamente, una distribución Normal:

$$\bar{\zeta} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Si tipificamos esta variable, obtendremos otra variable con distribución $N(0,1)$, que nos permitirá realizar inferencias sobre μ, cuando la desviación típica poblacional, σ, **es conocida**.

Es decir:

> Si la población sigue una distribución Normal en la variable considerada o el tamaño de las muestras es suficientemente grande (n>30) entonces, para realizar inferencias sobre μ, cuando **la desviación típica poblacional, σ, es conocida**, usaremos el siguiente estadístico:
>
> $$\frac{\bar{\zeta} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

En la práctica, la varianza poblacional no es conocida (si no conocemos la media, lo más probable es que tampoco conozcamos la varianza).

Entonces el estadístico anterior no nos sirve ya que depende de un parámetro desconocido. Para resolver el problema usaremos el siguiente resultado que utiliza la cuasidesviación

típica muestral en lugar de la desviación típica poblacional:

Si se toman muestras de tamaño n de una población que, para la variable considerada, sigue una distribución Normal de media μ (que queremos estimar), y **la desviación típica poblacional, σ, es desconocida**, entonces:

$$\frac{\bar{\zeta} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

(s es la cuasidesviación típica muestral).

6.2. Intervalo de confianza para la media

Hemos visto cómo se distribuye la media muestral, pero no olvidemos que a nosotros lo que nos interesa es poder **hacer una estimación de la media poblacional a partir de los resultados de una muestra**.

Lo que vamos a hacer es, partiendo de los resultados obtenidos para la muestra, construir un intervalo en el que «confiamos» que se encuentre la media poblacional.

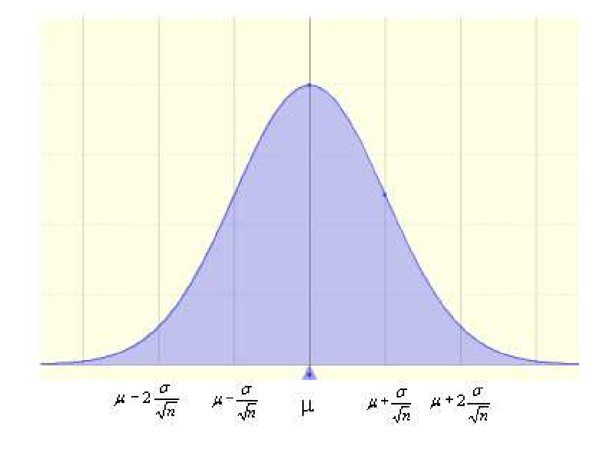
Llamaremos **nivel de confianza** al porcentaje de confianza que tenemos al hacer la estimación (también se puede expresar en términos de probabilidad como $1 - \alpha$), o bien, podemos hablar también del **nivel de significación**, α, que no es otra cosa que la probabilidad de error que estamos dispuestos a asumir en la estimación.

Estos dos conceptos son complementarios: Si estamos dispuestos a asumir una probabilidad de error de $\alpha = 0.05$ (5 % de error), entonces, nuestro nivel de confianza será del 95 % (ó 0.95 en términos de probabilidad).

Por otra parte, queda claro que cuanto mayor sea el error admitido, menor será el nivel de confianza.

La distribución de la media muestral para poblaciones normales o muestras grandes, con varianza conocida es:

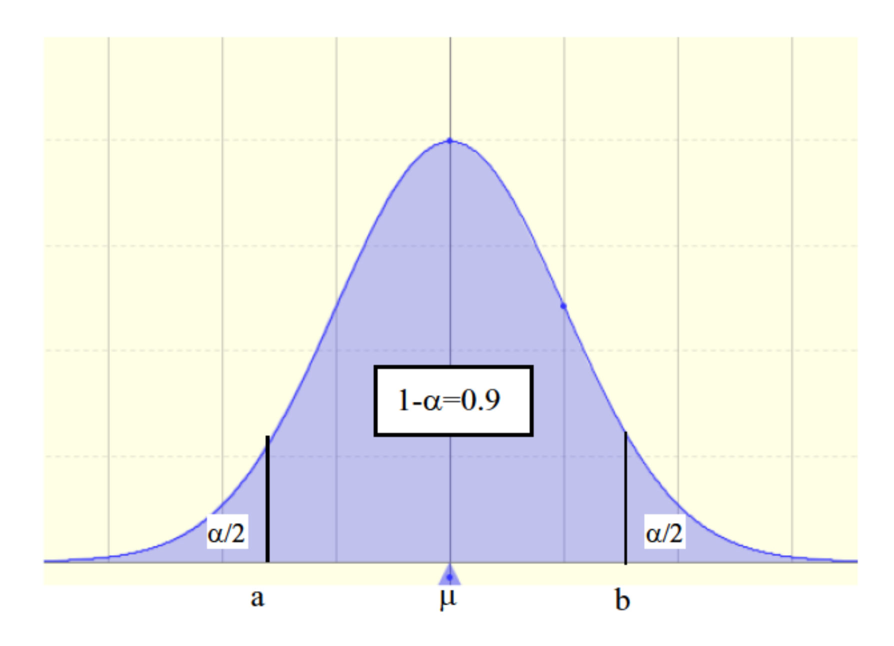
$$\bar{\zeta} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



Buscamos un intervalo, en torno a la media, que encierre una probabilidad del 90 % (por ejemplo).

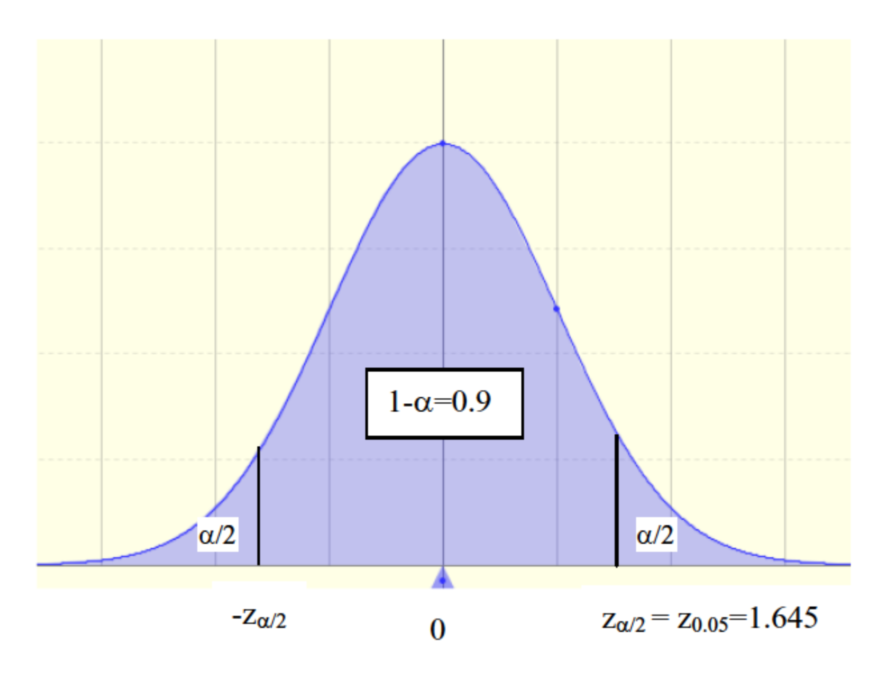
Este intervalo será aquel que cumpla que:

$$P\{|\bar{\zeta} - \mu| < x\} = 0.9, \text{ cuando } \bar{\zeta} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$



Para establecer los límites de este intervalo, tenemos que calcular el valor, para una $N(\mu, \frac{\sigma}{\sqrt{n}})$, que deja a su derecha una probabilidad igual a $\alpha/2$. Para hacerlo primero tipificamos, con lo que tenemos que:

$$Z = \frac{\bar{\zeta} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$



Entonces, si para la $Z = \frac{\bar{\zeta}-\mu}{\sigma/\sqrt{n}}$, el intervalo de confianza, con un nivel de confianza 1-α, es $(-z_{\alpha/2}, z_{\alpha/2})$, esto significa que:

$$1-\alpha = P\left\{-z_{\alpha/2} < \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\} = P\left\{\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\}$$

Por lo tanto:

Cuando trabajamos con poblaciones en las que la variable sigue una distribución **Normal**, o con **muestras grandes**, y además **la varianza poblacional es conocida**, el intervalo de confianza para la media, μ, con un nivel de confianza $1-\alpha$ es:

$$IC(\mu) = \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

donde, $z_{\alpha/2}$ es el valor de la $N(0,1)$, que deja a su derecha una probabilidad igual a $\alpha/2$.

Cuando la varianza poblacional no es conocida, no podemos calcular este intervalo. En ese caso usaremos el siguiente resultado que se obtiene siguiendo un razonamiento análogo para una t de Student:

Cuando trabajamos con poblaciones en las que la variable sigue una distribución **Normal**, y **la varianza poblacional es desconocida**, el intervalo de confianza para la media, μ, con un nivel de confianza $1-\alpha$ es:

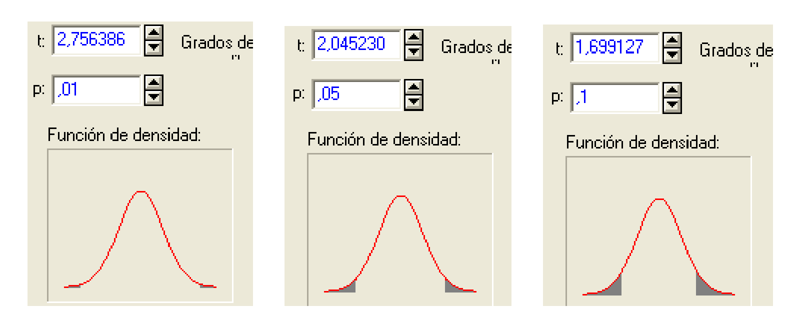
$$IC(\mu) = \left(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right)$$

donde, $t_{n-1,\alpha/2}$ es el valor de la t de Student con n-1 grados de libertad, que deja a su derecha una probabilidad igual a $\alpha/2$.

De ahora en adelante trabajaremos en este supuesto (varianza poblacional desconocida).

Aunque el error que estamos dispuestos a admitir puede cambiar, los valores más habituales son: 1 %, 5 % y 10 % (α=0.01, 0.05 y 0.1 respectivamente).

Por ejemplo, para una t de Student con 29 grados de libertad, estas situaciones son:



(donde t es el valor de $t_{n-1,\alpha/2} = t_{29,\alpha/2}$ y $p = \alpha$)

Una vez que hemos establecido el error máximo que estamos dispuestos a admitir en nuestra estimación, podemos establecer que la media poblacional se encuentra en el intervalo: $IC = \bar{x} \pm EM$, donde EM es el error muestral.

(el error muestral es la desviación respecto al parámetro)

En el ejemplo de las medias: para la última muestra de tamaño 30, teníamos que la media observada (media de la muestra) era $\bar{x} = 177.18$, con una cuasidesviación típica muestral de $s = 11.01$.

Entonces, el intervalo de confianza para la media poblacional con un nivel de confianza del 90 % será:

α**=0.1:** $IC(\mu) = 177.18 \pm 1.699127\frac{11.01}{\sqrt{30}} = 177.18 \pm 3.415486$

Es decir que, estimamos que la media poblacional se encuentra dentro del intervalo (173.7645, 180.5955), con un nivel de confianza del 90 %.

Los intervalos de confianza para la media con los otros niveles de significación más habituales son:

α**=0.01:** $IC(\mu) = 177.18 \pm 2.756386\frac{11.01}{\sqrt{30}} = 177.18 \pm 5.540727$

Es decir que, estimamos que la media poblacional se encuentra dentro del intervalo (171.6393, 182.7207), con un nivel de confianza del 99 %.

α**=0.05:** $IC(\mu) = 177.18 \pm 2.04523\frac{11.01}{\sqrt{30}} = 177.18 \pm 4.111202$

Es decir que, estimamos que la media poblacional se encuentra dentro del intervalo (173.0688, 181.2912), con un nivel de confianza del 95 %.

6.3. Contraste de hipótesis

En muchas ocasiones, el objetivo de nuestro análisis será corroborar empíricamente alguna hipótesis inicial sobre la población objeto de estudio. Muestra de ello pueden ser las siguientes situaciones:

1. Las especificaciones del fabricante indican que la vida media de una batería es de 4 años. Una organización de consumidores mantiene que la vida media de la batería es sensiblemente menor, y para comprobarlo experimentalmente, realizará un seguimiento sobre 40 usuarios de este tipo de baterías.

2. El Ministerio de Cultura de un país sostiene que el 60 % de los votantes apoyaría un incremento en el presupuesto de este ministerio, pero el gobierno no está dispuesto a modificar dicho presupuesto salvo que esa afirmación pueda ser corroborada científicamente. Con tal objetivo, el ministerio de cultura opta por hacer una encuesta a 2000 personas.

3. Un centro de investigación afirma que dispone de una vacuna contra la malaria más eficiente que la desarrollada por el Dr. Patarroyo. Esta vacuna fue probada sobre

38 voluntarios del Cuerpo de Paz que fueron a un país tropical en el que estaban especialmente expuestos a la enfermedad. A la mitad se les inoculó la vacuna del Dr. Patarroyo y a la otra mitad la nueva vacuna. De los que recibieron la nueva vacuna 15 se libraron de contraer la malaria, mientras que de los que recibieron la vacuna del Dr. Patarroyo, solamente 11.

Podemos observar que en todos los casos, hay una afirmación sobre los parámetros poblacionales y se toma una muestra para, con los resultados obtenidos para la misma, avalar o rechazar dicha afirmación.

En esencia éste es el planteamiento general de lo que en Inferencia Estadística se conoce como **pruebas o contrastes de hipótesis**.

- Se formula una hipótesis sobre la población.
- Se experimenta (la propia hipótesis nos sugiere cómo realizar el muestreo).
- Se decide si los resultados obtenidos para la muestra apoyan estadísticamente la hipótesis de partida.

Dado que nos movemos en condiciones de incertidumbre, esta última decisión se deberá tomar en términos probabilísticos, es decir, si los resultados obtenidos para la muestra tienen una alta probabilidad cuando la suposición de partida es cierta, entonces no tenemos evidencia en contra de dicha suposición (aceptamos la hipótesis de partida). Pero si los resultados obtenidos para la muestra son poco probables cuando suponemos que la hipótesis de partida es cierta, entonces, esto nos lleva a rechazar dicha hipótesis.

Veamos cómo desarrollar todo esto. En primer lugar vamos a definir una serie de términos:

Hipótesis nula (H_0) es la hipótesis que queremos contrastar. Es la hipótesis que el experimentador asume como correcta.

Hipótesis alternativa (H_1) es la negación de la hipótesis nula (es lo que aceptamos cuando rechazamos la hipótesis nula)

Estadístico de contraste (o medida de discrepancia) es cualquier función de los datos muestrales y del parámetro especificado por la hipótesis nula, con distribución conocida cuando H_0 es cierta.

Esta metodología, en la que la toma de decisiones está basada en los resultados obtenidos con una muestra, puede conducir a dos tipos de errores:

		Decisión	
		Aceptar H_0	Rechazar H_0
	H_0 verdadera	Correcto $(1-\alpha)$	Error de tipo I (α)
Realidad	H_0 falsa	Error de tipo II (β)	Correcto $(1-\beta)$

A la probabilidad de rechazar la hipótesis nula cuando es falsa $(1-\beta)$ se le llama potencia del contraste.

A la probabilidad de rechazar la hipótesis nula cuando es verdadera (α) se le llama nivel de significación.

Para construir y resolver un contraste de hipótesis, se siguen los pasos siguientes:

1. **Enunciar la hipótesis nula (H_0) y la hipótesis alternativa (H_1).**

   Ambas hipótesis deben ser excluyentes. La hipótesis nula es la que se considera como cierta. La hipótesis alternativa es la que aceptaremos solo si la muestra nos proporciona «suficiente evidencia en contra» de la hipótesis nula.

   Dependiendo de la formulación de la hipótesis alternativa, el contraste puede ser unilateral o bilateral.

2. **Determinar el nivel de significación.**

   Recordemos que el nivel de significación (α) es el nivel de error de tipo I que estamos dispuestos a aceptar. Los valores más habituales son: 0.01, 0.05 y 0.1.

   En muchas ocasiones se habla de nivel de confianza: 1-α (y se expresa en %).

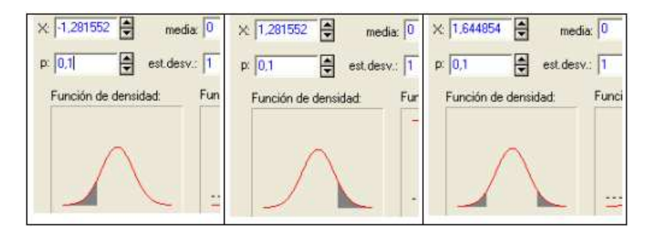
3. **Determinar el estadístico apropiado para la prueba y la zona de rechazo (llamada también región crítica).**

   Estos estadísticos vienen dados por la distribución muestral del estadístico objeto de estudio.

   Una vez que conocemos esa distribución, suponiendo que H_0 es cierta, tenemos que determinar la *zona de rechazo de la hipótesis nula.*

   Esta zona es el conjunto de valores posibles del estadístico que son tan extremos que la probabilidad de que ocurran, cuando H_0 es cierta, es muy pequeña (menor que α).

   En el caso de la $N(0,1)$:

   (De forma análoga se obtiene con otras distribuciones)

   La región que no es zona de rechazo se llama: *región de aceptación.*

4. **Calcular el estadístico.**

   Con los datos observados de la muestra y suponiendo que la hipótesis nula, H_0, es cierta, calculamos el estadístico y la probabilidad de encontrar un valor más alejado del parámetro que el que hemos calculado (p-**valor**).

5. **Tomar la decisión e interpretarla.**

   Aceptaremos H_0 si el p-valor es mayor que el nivel de significación (α).

   Si el p-valor es mayor que el nivel de significación esto es equivalente a decir que el valor del estadístico que hemos calculado está en la región de aceptación de H_0.

   ¡Ojo!, no es suficiente con decir «**se acepta** H_0», hay que explicar lo que significa aceptar o rechazar la hipótesis nula para el nivel de significación considerado.

6.4. Contraste de hipótesis para la media

Queremos contrastar que la media poblacional toma el valor μ_0.

Para ello, tomamos una muestra, y planteamos el siguiente contraste (bilateral):

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \\ n.\ significación : \alpha \end{cases}$$

α/2

Si los resultados muestrales no nos proporcionan evidencia en contra de la hipótesis nula, aceptaremos H_0, y en caso contrario la rechazaremos.

¿Cómo comprobamos esta evidencia?

Conocemos la distribución de la media muestral cuando la distribución de la variable poblacional es Normal:

$$\frac{\bar{\zeta} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Con los datos de la muestra (como suponemos que H_0 es cierta), calculamos el valor de prueba:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \text{ y su } p\text{-valor: } p = P\{|T| > t, \text{ dado que } T \sim t_{n-1}\}$$

Entonces, aceptaremos H_0 si $p > \alpha$ (no hay evidencia en contra de la hipótesis nula).

En algunos casos, lo que nos plantearemos no es un valor concreto del parámetro sino si el parámetro toma un valor mayor o menor que un valor dado (Ej.: vida media de una pila mayor que 3 años).

En estos casos el procedimiento es análogo. Solo **hay que tener cuidado al plantear la hipótesis nula** (es la que consideramos como cierta y queremos contrastar).

Podemos plantear dos situaciones:

$$\left| \begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \\ n.\ significación : \alpha \end{array} \right.$$

α

Entonces:

- $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- p-valor: $p = P\{T > t, \text{ dado que } T \sim t_{n-1}\}$
- Aceptaremos H_0 si $p > \alpha$

O bien:

$$\left| \begin{array}{l} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \\ n.\ significación : \alpha \end{array} \right.$$

α

Entonces:

- $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- p-valor: $p = P\{T < t, \text{ dado que } T \sim t_{n-1}\}$
- Aceptaremos H_0 si $p > \alpha$

Nota: en muchas ocasiones los programas de tratamiento estadístico solo hacen contrastes bilaterales, en estos casos, si usamos el p-valor para aceptar o rechazar una hipótesis de un contraste unilateral, consideraremos que:

- Si el signo de t es el contrario al de la zona de rechazo: Se acepta H_0 , ya que hay una gran evidencia a su favor (estamos en la cola contraria).
- Si t tiene el signo de la zona de rechazo: Se acepta H_0 si $p/2 > \alpha$

Un estudio análogo al realizado con la media muestral se puede realizar para la proporción muestral.

Relación entre el contraste de hipótesis y el intervalo de confianza

Los contrastes de hipótesis y los intervalos de confianza tienen una estrecha relación.

Para un determinado nivel de confianza, esta relación podríamos expresarla diciendo que «El intervalo de confianza está formado por todos los valores del parámetro que se aceptarían en un contraste de hipótesis bilateral».

Dicho de otra forma, cualquier valor del parámetro que se encuentre dentro del intervalo de confianza dará lugar a un valor del estadístico que se encuentra dentro de la región de aceptación en el contraste de hipótesis bilateral correspondiente.

Por lo tanto, en el caso de la media poblacional tenemos que:

Si con los datos de la muestra, $\mu_0 \in IC_{1-\alpha}(\mu)$ entonces, si planteamos el contraste de hipótesis:

$H_0 : \mu = \mu_0$
$H_1 : \mu \neq \mu_0$
con nivel de significación α

y lo queremos resolver apoyándonos en los datos de la misma muestra, el resultado será que debemos aceptar la hipótesis nula. Es decir, que el valor μ_0 es un valor aceptable para la media poblacional, para el nivel de significación dado, α.

Si por el contrario $\mu_0 \notin IC_{1-\alpha}(\mu)$, entonces, deberemos rechazar la hipótesis nula. Es decir que el valor μ_0 no es un valor aceptable para la media poblacional, para el nivel de significación dado, α.

6.5. Distribución de la proporción muestral

En muchas ocasiones nos interesará estimar no el valor medio de un conjunto de observaciones sino la proporción de veces que ocurre un determinado fenómeno (por ejemplo: la proporción de votantes a un partido político).

Es decir, queremos estimar el valor de la proporción poblacional (**p**), a partir de los resultados obtenidos con una muestra de tamaño n.

Siguiendo un razonamiento análogo al utilizado para determinar la distribución de la media muestral (tomando muchas muestras de tamaño n, calculando la proporción para cada una de ellas y estudiando la distribución de la variable *proporción muestral*, $\hat{p}$), llegamos a que:

Cuando n es grande ($n > 30$), la distribución de la proporción muestral es una Normal:

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

o lo que es equivalente:

$$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

6.5.1. Intervalo de confianza para una proporción

Haciendo lo mismo que en el caso de la media poblacional, para calcular el intervalo de confianza de la proporción poblacional, con un nivel confianza $1-\alpha$ (o un nivel de significación α), tenemos que:

$$IC(p) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

Este intervalo no lo podemos calcular ya que depende del parámetro p que queremos estimar.

Lo que hacemos es utilizar para el cálculo, la proporción muestral en lugar de la proporción poblacional. Entonces:

Cuando n es suficientemente grande ($n > 30$), el intervalo de confianza para la proporción poblacional, con un nivel de significación α, es:

$$IC(p) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

NOTACIÓN: $z_{\alpha/2}$ es el valor de una $N(0,1)$ que deja a la derecha una probabilidad $\alpha/2$

También nos podemos poner en la peor situación posible y determinar el intervalo de confianza más grande posible, para un nivel de confianza 1-α, que es el que se obtiene cuando p=1/2.

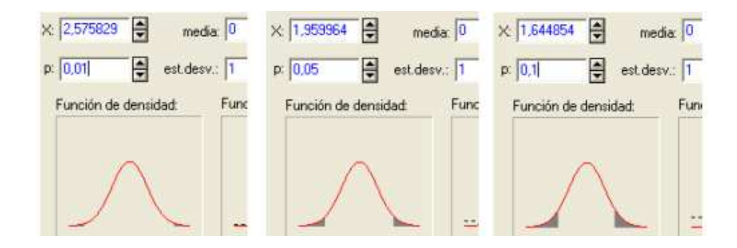
En esa situación:

Cuando n es suficientemente grande ($n > 30$), el **intervalo de confianza más grande posible** para la proporción poblacional, con un nivel de significación α, es:

$$IC(p) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{4n}} = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}}\right)$$

NOTACIÓN: $z_{\alpha/2}$ es el valor de una $N(0,1)$ que deja a la derecha una probabilidad $\alpha/2$

Para los niveles de significación más habituales, los valores de $z_{\alpha/2}$, son:



Veamos un ejemplo:

Si basándonos en una muestra (altura de 30 estudiantes) queremos determinar la proporción de estudiantes universitarios con una altura superior a 180 cm, deberíamos hacer lo siguiente:

1. Observamos lo que ocurre en la muestra: 5 de los estudiantes tienen una altura superior a los 180 cm.

   Esto significa que la proporción muestral es: $\hat{p} = \frac{5}{30} = 0.1\hat{6}$

2. Determinamos el nivel de error que estamos dispuestos a aceptar y construimos el intervalo de confianza correspondiente. Así:

   - Si $\alpha = 0.1$:

$$IC(p) = \frac{5}{30} \pm 1.644854\sqrt{\frac{\frac{5}{30}\frac{25}{30}}{30}}$$

$$IC(p) = (0.05475, 0.27858)$$

3. Interpretamos el resultado:

   Basándonos en los resultados de la muestra, estimamos con un nivel de confianza del 90 %, que la proporción poblacional de estudiantes con una altura mayor que 180 cm se encuentra dentro del intervalo: (0.05475, 0.27858) (o lo que es lo mismo, entre el 5.475 % y el 27.858 %).

Los intervalos de confianza para otros niveles de significación son:

Si $\alpha = 0.01$:

$$IC(p) = \frac{5}{30} \pm 2.575829\sqrt{\frac{\frac{5}{30}\frac{25}{30}}{30}}$$

Es decir que, basándonos en los resultados de la muestra, estimamos con un nivel de confianza del 99 %, que la proporción poblacional de estudiantes con una altura mayor que 180 cm se encuentra dentro del intervalo: (-0.008596, 0.341930).

Este resultado teóricamente está bien, pero como la proporción no puede ser negativa, podemos utilizar el intervalo: (0, 0.34193).

Es decir que estimamos, con un nivel de confianza del 99 %, que la proporción poblacional de estudiantes con una altura mayor que 180 cm es menor que el 34.193 %.

Si $\alpha = 0.05$:

$$IC(p) = \frac{5}{30} \pm 1.959964\sqrt{\frac{\frac{5}{30}\frac{25}{30}}{30}}$$

Es decir que, basándonos en los resultados de la muestra, estimamos con un nivel de confianza del 95 %, que la proporción poblacional de estudiantes con una altura mayor que 180 cm se encuentra dentro del intervalo: (0.03331, 0.30003) (o lo que es lo mismo, entre el 3.331 % y el 30.003 %).

Nota: la mayoría de los programas de tratamiento estadístico no calculan los intervalos de confianza para una proporción, en ese caso, tendremos que calcularlos nosotros.

6.5.2. Contraste de hipótesis para una proporción

Nota: *en este apartado*, para no confundir la proporción poblacional ,p, con el p-valor, a este último lo llamaremos siempre p-valor.

Queremos contrastar que la proporción poblacional toma el valor: p_0.

Para ello, tomamos una muestra, y planteamos el siguiente contraste (bilateral):

$H_0 : p = p_0$
$H_1 : p \neq p_0$
n. significación : α

α/2

Si los resultados muestrales no nos proporcionan evidencia en contra de la hipótesis nula, aceptaremos H_0, y en caso contrario la rechazaremos.

¿Cómo comprobamos esta evidencia?

Conocemos la distribución de la proporción muestral cuando n es suficientemente grande:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Entonces, para nuestra muestra y suponiendo que H_0 es cierta, calculamos:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \text{ y su } p\text{-valor es: } p\text{-valor} = P\{|Z| > z, \text{ donde } Z \sim N(0,1)\}$$

($\hat{p}$ es la proporción muestral)

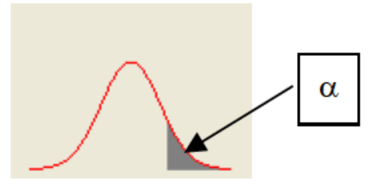
Entonces, aceptaremos H_0 si p-valor$> \alpha$ (no hay evidencia en contra de la hipótesis nula).

En algunos casos, lo que nos plantearemos no es un valor concreto del parámetro sino si el parámetro toma un valor mayor o menor que un valor dado (Ej.: proporción de votantes menor que el 60 %).

En estos casos el procedimiento es análogo. Solo hay que tener cuidado al plantear la hipótesis nula (es la que consideramos como cierta y queremos contrastar).

Podemos plantear dos situaciones:

$$\left| \begin{array}{l} H_0 : p \leq p_0 \\ H_1 : p > p_0 \\ n.\ significación : \alpha \end{array} \right.$$



Entonces:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \text{ y su } p\text{-valor es: } p\text{-valor} = P\{Z > z, \text{ donde } Z \sim N(0,1)\}$$

Aceptaremos H_0 si p-valor$> \alpha$

O bien,

$$\left| \begin{array}{l} H_0 : p \geq p_0 \\ H_1 : p < p_0 \\ n.\ significación : \alpha \end{array} \right.$$

Entonces:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \text{ y su } p\text{-valor es: } p\text{-valor} = P\{Z < z, \text{ donde } Z \sim N(0,1)\}$$

Aceptaremos H_0 si p-valor$> \alpha$

6.6. Contraste de igualdad (o diferencia) de medias

Los contrastes anteriores se referían al valor del parámetro poblacional, es decir, se contrasta si se puede aceptar o no que el parámetro poblacional toma un determinado valor, pero en muchas ocasiones lo que nos interesa es contrastar si dos muestras provienen de poblaciones en las que la variable tiene la misma media (contrastaremos la igualdad de medias).

En este caso vamos a proceder igual que en los casos anteriores, aunque no justificaremos los estadísticos de contraste.

Estos estadísticos los veremos solamente a título informativo ya que los

problemas correspondientes los resolveremos con el ordenador.

Queremos contrastar si dos poblaciones tienen, para la variable objeto de estudio, la misma media. O dicho de otra forma, si existen diferencias, estadísticamente significativas entre dos medias muestrales.

Para ello, tomaremos dos muestras (una de cada población), que no tienen por qué ser del mismo tamaño: $\zeta_1, ..., \zeta_n$ y $\xi_1, ..., \xi_m$. Estas muestras deben ser aleatorias.

Distinguiremos las siguientes situaciones:

- **Si las muestras son pareadas**.

  En este caso las muestras sí que tienen que ser del mismo tamaño. Este sería el caso de dos características estudiadas para el mismo individuo (por ejemplo, tensión antes y después de un tratamiento). En la práctica lo que se hace, en lugar de contrastar si las medias son iguales, es contrastar si la variable diferencia de estas dos variables tiene media cero. Por lo tanto este caso se reduce a un contraste de una media (con todas sus condiciones de validez).

- **Si las muestras son independientes la una de la otra**.

  Este caso corresponde a la comparación de las medias de una misma variable para dos grupos independientes de casos (estudiar si se puede aceptar que existen diferencias estadísticamente significativas entre los salarios medios de dos categorías de empleados -o entre los tiempos medios de fabricación de una pieza en el turno de mañana y en el de tarde -).

El procedimiento para realizar los contrastes de igualdad de medias es el mismo que en los dos casos anteriores. Las distribuciones utilizadas para establecer el estadístico de contraste dependen de las características de las variables observadas:

- Poblaciones Normales o poblaciones cualesquiera cuando tenemos muestras grandes y las varianzas poblacionales son conocidas:

$$\frac{\bar{\zeta}_1 - \bar{\zeta}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

- Poblaciones Normales y las varianzas poblacionales son desconocidas pero podemos aceptar que son iguales:

$$\frac{\bar{\zeta}_1 - \bar{\zeta}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

donde: $S_p = \sqrt{\dfrac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}}$, (s_1^2 y s_2^2 son las cuasivarianzas muestrales).

6.7. Contraste de igualdad (o diferencia) de proporciones

En el caso del contraste para la diferencia de proporciones, seguiremos el mismo procedimiento que en los casos anteriores.

El estadístico de contraste que utilizaremos se basa en el siguiente resultado:

Si tenemos **muestras grandes** se cumple que:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}} \sim N(0,1)$$

Como las proporciones poblacionales p_1 y p_2 son desconocidas, para calcular el estadístico de contraste se suelen utilizar, en el denominador, las proporciones muestrales, $\hat{p}_1$ y $\hat{p}_2$.

En muchas ocasiones, como la hipótesis nula supone que las proporciones poblacionales son iguales ($H_0 : p_1 = p_2 = p$), se utiliza una única estimación de la proporción poblacional común:

$$\hat{p} = \frac{n\hat{p}_1 + m\hat{p}_2}{n+m}$$

En este caso el estadístico de contraste será:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n} + \frac{1}{m}\right)}} \quad \text{donde } \hat{p} = \frac{n\hat{p}_1 + m\hat{p}_2}{n+m}$$

6.8. Ejemplos resueltos

1. Se desea investigar la superficie (en m^2) de los pisos en venta en el mercado inmobiliario de nuestra ciudad. Para ello se cuenta con la información proporcionada por una muestra aleatoria de 31 viviendas:

SUPERFICIE (en m^2)									
85.8	49.2	65.7	58.3	65.2	52.9	87.2	90.3	51.6	69.4
75.9	70.3	76.0	71.9	65.2	82.3	70.9	70.8	74.6	52.1
34.1	81.9	86.1	76.1	55.4	63.2	105.1	59.9	62.7	85.9
84.5									

Describiendo las hipótesis adecuadas, contesta a las siguientes preguntas:

a) Determina los **intervalos de confianza** para la media de la superficie de los pisos en el mercado inmobiliario de nuestra ciudad con niveles de confianza del 90 % y del 99 %.

b) A partir del resultado obtenido en el apartado anterior y, **sin realizar ningún cálculo adicional**, ¿cuál es nuestra conclusión ante las hipótesis:
H_0 : la superficie media de los pisos ofertados es igual a 75 m^2
H_1 : la superficie media de los pisos ofertados no es igual a 75 m^2
con niveles de significación α =0.1, 0.01 y 0.05?

c) Responde a todas las preguntas del apartado anterior (haciendo los cálculos que sean necesarios).

d) Un directivo inmobiliario afirma que el porcentaje de pisos de menos de 60 m^2 no supera el 15 %. ¿Se sostiene su afirmación con un nivel de significación de 0.05?

e) Se realiza un estudio similar en otra ciudad y, una muestra aleatoria de 21 pisos, da como resultado (llamemos Y a la superficie de los pisos en esta nueva ciudad):

$$\bar{Y} = 85\ m^2\ ,\ S_Y^2 = 200\ m^4$$

Con un nivel de significación 0.05, ¿puede descartarse que la media de la superficie de los pisos en el mercado inmobiliario de nuestra ciudad sea similar a la de esta otra?

Solución:

Nos dan una muestra de la superficie en m^2, X, de 31 viviendas (n=31). Para esta muestra podemos calcular la media y la cuasidesviación típica:

$$\bar{x} = 70.871 \text{ y } S = 14.86032$$

a) Determina los intervalos de confianza para la media de la superficie de los pisos en el mercado inmobiliario de nuestra ciudad con niveles de confianza del 90 % y del 99 %.

Nos piden IC(μ), para $1 - \alpha = 0.9$ y para $1 - \alpha = 0.99$

Como la varianza poblacional es desconocida, **suponiendo que la superficie de los pisos sigue una distribución Normal**, se verifica que:

$$\frac{\bar{\zeta} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Entonces: $\text{IC}_{1-\alpha}(\mu) = \bar{x} \pm t_{n-1,\alpha/2}\dfrac{s}{\sqrt{n}}$

$\text{IC}_{0.9}(\mu) = 70.33871 \pm t_{30,0.05}\dfrac{14.86032}{\sqrt{31}} = 70.33871 \pm 1.6973 \times 2.66899$

$\text{IC}_{0.9}(\mu) = 70.33871 \pm 4.53008 = (65.80863, 74.86879)$

La superficie media de los pisos se encuentra entre 65.81 m^2 y 74.87 m^2, con un nivel de confianza del 90 %.

Análogamente:

$\text{IC}_{0.99}(\mu) = 70.33871 \pm t_{30,0.005} \times 2.66899 = 70.33871 \pm 2.75 \times 2.66899$

$\text{IC}_{0.99}(\mu) = 70.33871 \pm 7.33972 = (62.99899, 77.67843)$

La superficie media de los pisos se encuentra entre 63 m^2 y 77.68 m^2, con un nivel de confianza del 99 %.

b) A partir del resultado obtenido en el apartado anterior y, **sin realizar ningún cálculo adicional**, ¿cuál es nuestra conclusión ante las hipótesis:

$\left| \begin{array}{l} H_0 : \mu = 75 \\ H_1 : \mu \neq 75 \end{array} \right.$ con niveles de significación α =0.1, 0.01 y 0.05?

Para responder a esta pregunta utilizaremos los intervalos de confianza calculados en el apartado anterior, por lo tanto, las condiciones de validez son las mismas que en dicho apartado: como no conocemos la varianza poblacional, para poder hacer inferencias sobre la media poblacional **necesitamos que la variable superficie siga una distribución Normal**.

Para $\alpha = 0.1$: esto significa que $1 - \alpha = 0.9$
$75 \notin \mathrm{IC}_{0.9}(\mu)$, por lo tanto, **rechazamos** H_0.
Es decir: para un nivel de significación del 10 %, no podemos aceptar que la superficie media de las viviendas en venta sea de 75 m^2.

Para $\alpha = 0.01$: esto significa que $1 - \alpha = 0.99$
$75 \in \mathrm{IC}_{0.99}(\mu)$, por lo tanto, **aceptamos** H_0.
Es decir: para un nivel de significación del 1 %, aceptamos que la superficie media de las viviendas en venta es de 75 m^2.

Para $\alpha = 0.05$: esto significa que $1 - \alpha = 0.95$
Sin hacer más cálculos **no podemos decir nada**
A partir de los resultados anteriores «da la sensación» de que con este nivel de significación se aceptaría la hipótesis nula (ya que 75 está muy cerca del extremo del intervalo de confianza para $\alpha = 0.1$) y para $\alpha = 0.05$ el intervalo es mayor). Pero en realidad no podemos afirmar que vaya a estar dentro.

c) Responde a todas las preguntas del apartado anterior (haciendo los cálculos que sean necesarios).

Resolvemos el contraste de hipótesis en general y luego respondemos a las cuestiones en función del nivel de significación.

Queremos resolver el contraste: $\left| \begin{array}{l} H_0 : \mu = 75 \\ H_1 : \mu \neq 75 \end{array} \right.$ para distintos niveles de significación.

Como la varianza poblacional es desconocida, **suponiendo que la superficie de los pisos sigue una distribución Normal**, se verifica que:

$$\frac{\bar{\zeta} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Entonces, si H_0 es cierta, el valor del estadístico $\dfrac{\bar{x} - 75}{s/\sqrt{n}}$ es un valor de una t_{n-1}

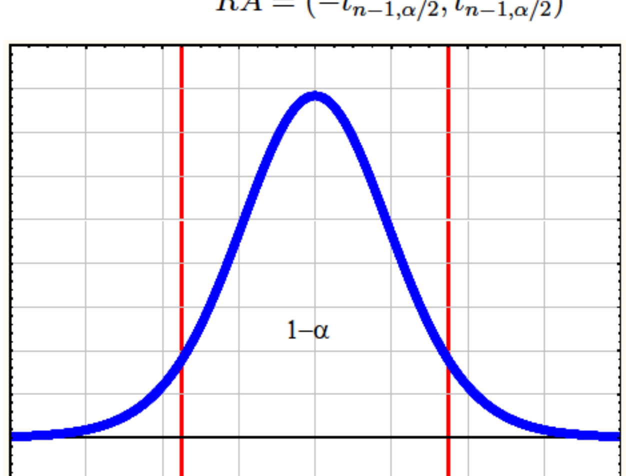
Es decir que, en nuestro caso

$$t = \frac{70.33871 - 75}{\frac{14.86032}{\sqrt{31}}} = \frac{-4.66129}{2.66899} = -1.74646$$

es un valor de una t_{30}

Para saber si se acepta o no la hipótesis nula, debemos determinar si el estadístico calculado se encuentra o no en la región de aceptación y dicha región es el intervalo:

$$RA = (-t_{n-1,\alpha/2}, t_{n-1,\alpha/2})$$



Entonces:

Para $\alpha = 0.1$:

Buscamos en las tablas: $t_{30,\alpha/2} = t_{30,0.05} = \mathbf{1.6973}$,

A continuación comprobamos si nuestro estadístico pertenece a la región de aceptación:

$$-1.74646 \notin (-1.6973, 1.6973)$$

Por lo tanto, rechazamos la hipótesis nula.

Es decir que, para un nivel de significación de 0.1 no se puede aceptar que la superficie media de las viviendas sea de 75 m^2.

Para $\alpha = 0.01$:

Buscamos en las tablas: $t_{30,\alpha/2} = t_{30,0.005} = \mathbf{2.75}$,

Comprobamos si nuestro estadístico pertenece a la región de aceptación:

$$-1.74646 \in (-2.75, 2.75)$$

Por lo tanto, aceptamos la hipótesis nula.

Es decir que, para un nivel de significación de 0.01 sí se puede aceptar que la superficie media de las viviendas sea de 75 m^2.

Por último, **para** $\alpha = 0.05$:

Buscamos en las tablas: $t_{30,\alpha/2} = t_{30,0.025} = \mathbf{2.0423}$,

Comprobamos si nuestro estadístico pertenece a la región de aceptación:

$$-1.74646 \in (-2.0423, 2.0423)$$

Por lo tanto, aceptamos la hipótesis nula.

Es decir que, para un nivel de significación de 0.05 sí se puede aceptar que la superficie media de las viviendas sea de 75 m^2.

Este apartado también se puede resolver **utilizando el p-valor**.
Una vez calculado el valor del estadístico: $t = -1.74646$, entonces:
p-valor$=P\{\mid T \mid > 1.74646$ siendo $T \sim t_{30}\}$
p-valor$=2 \times P\{T > 1.74646$ siendo $T \sim t_{30}\}$

Si usamos las tablas, teniendo en cuenta que $T \sim t_{30}$, las mejores aproximaciones que podemos hacer son :
p-valor$= 2 \times P\{T > 1.74646\} > 2 \times 0.025 = 0.05$
p-valor$= 2 \times P\{T > 1.74646\} < 2 \times 0.05 = 0.1$

Si usamos el ordenador obtenemos que el p-valor=0.090964

En cualquier caso, el razonamiento es el mismo:

- Si $\alpha = 0.1$, el p-valor es MENOR que α y rechazamos H_0.
- Si $\alpha = 0.05$, el p-valor es MAYOR que α y aceptamos H_0.
- Si $\alpha = 0.01$, el p-valor es MAYOR que α y aceptamos H_0.

Las respuestas son las mismas que hemos dado antes.

d) Un directivo inmobiliario afirma que el porcentaje de pisos de menos de 60 m^2 no supera el 15 %. ¿Se sostiene su afirmación con un nivel de significación de 0.05?

Lo que se plantea ahora es un contraste para una proporción.

Además este contraste es unilateral (la zona de rechazo es solo una de las colas).

Para poder hacer inferencia sobre proporciones y que el estadístico utilizado sea válido, necesitamos que el tamaño de la muestra sea suficientemente grande $n > 30$ y además que $np > 5$ y $n(1 - p) > 5$.

En este caso, como p=0.15 y n=31, no se cumplen las condiciones de validez porque, aunque n es suficientemente grande, $n > 30$, no se verifica otra de las condiciones $np = 0.15 \times 31 = 4.65 < 5$.

Resolvemos el problema suponiendo que se cumplen las condiciones de validez.

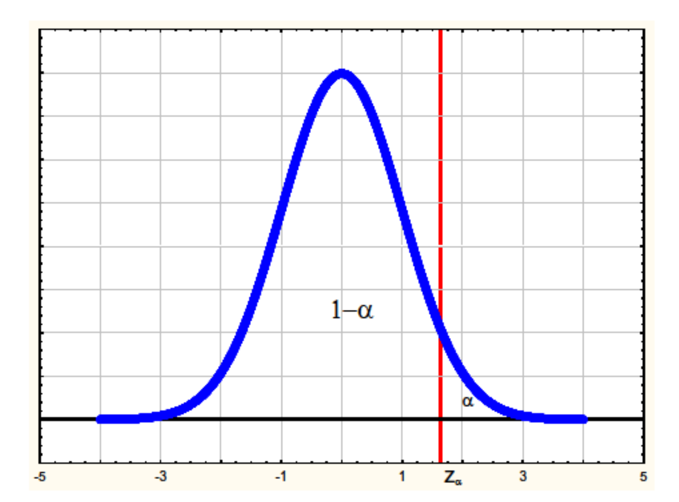
El contraste es:

$$\left| \begin{array}{l} H_0 : p \leq 0.15 \\ H_1 : p > 0.15 \end{array} \right. \quad \text{para un nivel de significación } \alpha = 0.05.$$

Si se cumplen las condiciones de validez sabemos que

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p}{n}}} \sim N(0, 1)$$

Gráficamente la situación es (la zona de rechazo está a la derecha):



Si H_0 es cierta, sabemos que el valor $z = \frac{\hat{p} - 0.15}{\sqrt{\frac{0.15(1-0.15)}{n}}}$ es un valor de una $N(0,1)$, y debemos determinar si dicho valor se encuentra en la región de aceptación o en la de rechazo para nuestro problema.

En este problema, como $\alpha = 0.05$, la zona de rechazo la determina el valor $z_\alpha = z_{0.05} = 1.645$

Calculamos el estadístico para la muestra:

$$z = \frac{\frac{8}{31} - 0.15}{\sqrt{\frac{0.15(1-0.15)}{31}}} = 1.76223$$

Como $1.76223 > 1.645$, esto significa que el estadístico se encuentra en la zona de rechazo. Por lo tanto: **rechazamos la hipótesis nula**.

Para un nivel de significación de 0.05 NO se sostiene la afirmación del directivo. Es decir, que debemos aceptar que la proporción de pisos con una superficie menor que 60 m^2 es mayor que el 15 %.

e) Se realiza un estudio similar en otra ciudad y, una muestra aleatoria de 21 pisos, da como resultado (llamemos Y a la superficie de los pisos en esta nueva ciudad):

$$\bar{Y} = 85\ m^2\ ,\ S_Y^2 = 200\ m^4$$

Con un nivel de significación 0.05, ¿puede descartarse que la media de la superficie de los pisos en el mercado inmobiliario de nuestra ciudad sea similar a la de esta otra?

Vamos a hacer un contraste de igualdad de medias para dos muestras independientes.

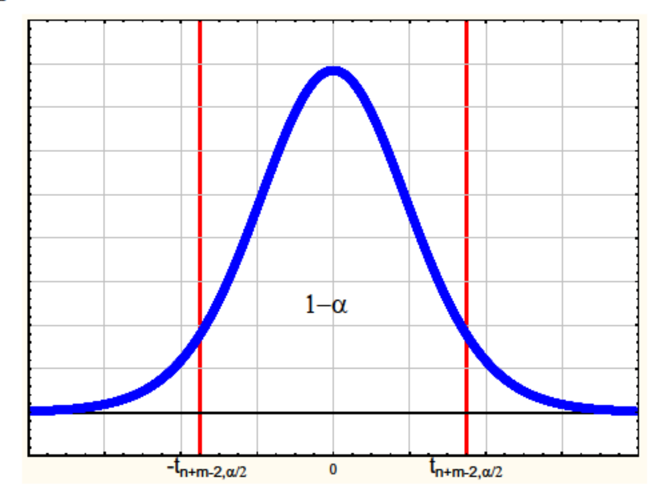
Como las varianzas poblacionales son desconocidas, para que los resultados del contraste sean válidos **necesitamos que la superficie de las viviendas, en ambas ciudades, siga una distribución Normal y que las varianzas de dichas superficies, aunque sean desconocidas, sean iguales.**

Si se cumplen las condiciones anteriores, sabemos que:

$$\frac{\bar{\zeta}_1 - \bar{\zeta}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2} \text{ donde } S_p = \sqrt{\frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}}$$

El contraste es:

$H_0 : \mu_X = \mu_Y$
$H_1 : \mu_X \neq \mu_Y$
$\alpha = 0.05$



Los límites de la región de aceptación los establece el valor:

$$t_{n+m-2,\alpha/2} = t_{31+21-2,0.025} = t_{50,0.025} = 2.0086$$

Región de aceptación: $RA = (-2.0086, 2.0086)$
Calculamos el estadístico para las muestras y determinamos si se encuentra o no en la región de aceptación:

$$t = \frac{70.33871 - 85}{\sqrt{\frac{30\times 220.829 + 20\times 200}{50}}\sqrt{\frac{1}{31} + \frac{1}{21}}} = \frac{-14.66129}{4.11677} = -3.56136$$

Como $t = -3.56136 < -2.0086$, está fuera de la región de aceptación, por lo tanto rechazamos que las medias son iguales para un nivel de significación del 5 %.
Respondemos a la pregunta:
SÍ, para un nivel de significación del 5 %, podemos descartar que la superficie media de los pisos en venta en nuestra ciudad es similar a los de esta otra.

2. El número de aciertos de 10 individuos en un determinado test psicotécnico, antes y después de echarse la siesta, fueron:

Antes (X)	232	249	246	243	213	215	246	283	247	244
Después (Y)	224	253	232	252	219	206	233	268	237	227

Describiendo las hipótesis adecuadas, contrasta con un nivel de significación α=0.05 la hipótesis nula de que tras una siestecita uno se encuentra «más despierto».

Solución:

Ese «más despierto» se traduce en que el número de aciertos después de la siesta es mayor que antes. Entonces, lo que debemos contrastar son las medias de 2 muestras pareadas (relacionadas).

Construimos una variable con las diferencias para cada caso ($D = Y - X$), y entonces contrastar la igualdad de medias en este problema, es equivalente a contrastar si la media de la variable diferencia es igual a cero:

(en nuestro caso el contraste será unilateral)

$\left| \begin{array}{l} H_0 : \mu_Y \geq \mu_X \\ H_1 : \mu_Y < \mu_X \\ \alpha = 0.05 \end{array} \right.$ es equivalente a $\left| \begin{array}{l} H_0 : \mu_D = \mu_{Y-X} \geq 0 \\ H_1 : \mu_D < 0 \\ \alpha = 0.05 \end{array} \right.$

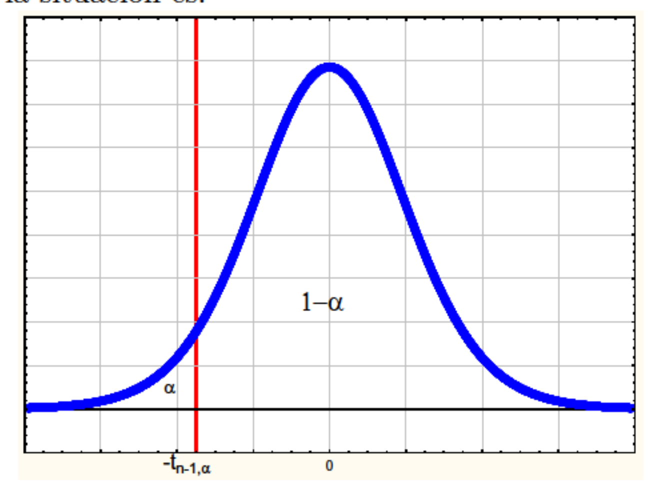
Antes (X)	232	249	246	243	213	215	246	283	247	244
Después (Y)	224	253	232	252	219	206	233	268	237	227
Diferencia (D)	-8	4	-14	9	6	-9	-13	-15	-10	-17

Para la variable D tenemos: $\bar{D} = -6.7$ y $s_D = 9.4757$

Entonces, como no conocemos la varianza poblacional de la diferencia, **si la variable diferencia sigue una distribución Normal**, sabemos que:

$$\frac{\bar{\zeta} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Gráficamente, la situación es:



El límite de la región de aceptación es: $-t_{n-1,\alpha} = -t_{9,0.05} = -1.8331$

Calculamos el valor del estadístico para ver si está en la región de aceptación:

$t = \dfrac{(-6.7) - 0}{9.4757/\sqrt{10}} = -0.2236$

El estadístico se encuentra dentro de la región de aceptación, lo que significa que aceptamos la hipótesis nula ($\mu_D \geq 0$).

Por lo tanto aceptamos que el número medio de aciertos después de la siesta es mayor que antes (no hay evidencia en contra), con un nivel de significación del 5 %.

También podíamos haberlo resuelto calculando el p-valor:

$$p = P\{T < -0.2236 \text{ , siendo } T \sim t_9\}$$

aunque las tablas no nos dan un valor exacto, para una t_9 sabemos que

$$p = P\{T < -0.2236\} = P\{T > 0.2236\} > 0.35 > \alpha = 0.05$$

Por lo tanto, como el p-valor$> \alpha$, aceptamos la hipótesis nula, es decir, aceptamos que el número medio de aciertos después de la siesta es mayor que antes (no hay evidencia en contra), con un nivel de significación del 5 %.

Resumen de Intervalos y Contrastes

Intervalo de confianza y contraste de hipótesis para la media:

Condiciones	Distribución	Intervalo de confianza	Contraste	Estadístico
Población Normal o población cualquiera con muestra grande ($n \geq 30$) y **varianza conocida**	$\frac{\bar{\zeta}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$	$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$
Población Normal y **varianza desconocida**	$\frac{\bar{\zeta}-\mu}{s/\sqrt{n}} \sim t_{n-1}$	$\left(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right)$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$t = \frac{\bar{x}-\mu_0}{\mathbf{s}/\sqrt{n}}$

Intervalo de confianza y contraste de hipótesis para la proporción:

Condiciones	Distribución	Intervalo de confianza	Contraste	Estadístico
Muestra grande ($n\hat{p} \geq 5$ y $n(1-\hat{p}) \geq 5$)	$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$	$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$	$H_0 : p = p_0$ $H_1 : p \neq p_0$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Intervalo de confianza y contraste de hipótesis para la diferencia de medias:

Condiciones	Distribución	Intervalo de confianza	Contraste	Estadístico
Poblaciones Normales o poblaciones cualesquiera con muestras grandes y **varianzas conocidas**	$\frac{\bar{\zeta}_1-\bar{\zeta}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \sim N(0,1)$	$\left(\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)$	$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 \neq d_0$	$z = \frac{\bar{x}-\bar{y}-d_0}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}}$
Poblaciones Normales y **varianzas desconocidas pero iguales**	$\frac{\bar{\zeta}_1-\bar{\zeta}_2-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$ donde $S_p = \sqrt{\frac{(n-1)s_1^2+(m-1)s_2^2}{n+m-2}}$	$\left(\bar{x} - \bar{y} - t_{n+m-2,\alpha/2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}, \bar{x} - \bar{y} + t_{n+m-2,\alpha/2}S_p\sqrt{\frac{1}{n} + \frac{1}{m}}\right)$	$H_0 : \mu_1 - \mu_2 = d_0$ $H_1 : \mu_1 - \mu_2 \neq d_0$	$t = \frac{\bar{x}-\bar{y}-d_0}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$

Intervalo de confianza y contraste de hipótesis para la diferencia de proporciones:

Condiciones	Distribución	Intervalo de confianza	Contraste	Estadístico
Muestras grandes $q = 1 - p$	$\frac{\hat{p}_1-\hat{p}_2-(p_1-p_2)}{\sqrt{\frac{p_1q_1}{n}+\frac{p_2q_2}{m}}} \sim N(0,1)$	$\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n} + \frac{\hat{p}_2\hat{q}_2}{m}}, \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n} + \frac{\hat{p}_2\hat{q}_2}{m}}\right)$	$H_0 : p_1 - p_2 = 0$ $H_1 : p_1 - p_2 \neq 0$	$z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n}+\frac{1}{m}\right)}}$ donde $\hat{p} = \frac{n\hat{p}_1+m\hat{p}_2}{n+m}$

Tema 7

Muestreo

En el tema anterior hablábamos de algunos de los aspectos más elementales de la Inferencia Estadística y dábamos por hecho que teníamos una muestra de tamaño n, con la que podíamos hacer inferencias, pero las cosas no son tan sencillas.

En primer lugar, una muestra debe ser un **subconjunto** de la población, pero no cualquier subconjunto, sino que debe ser **representativo** de la misma.

La forma en la que se elige este subconjunto influirá en los resultados que obtengamos. Hay toda una teoría de muestras desarrollada para determinar en cada caso cómo debe ser la muestra, en función del estudio que queramos realizar.

Este no es el momento ni el lugar para desarrollar esta teoría, pero sí que nos conviene conocer algunos de los tipos o técnicas más habituales de muestreo, ya que los necesitaremos para realizar cualquier estudio estadístico.

En segundo lugar, ¿cuál debe ser el **tamaño de la muestra**?, ¿es apropiado tomar 20 elementos?, ¿existen grandes diferencias entre tomar 50 o 100 elementos?, **¿en qué influye** el tamaño de la muestra?

Vamos a intentar dar una respuesta rápida y sencilla a estas dos cuestiones, aunque, como ya se ha dicho, el tema no es ni rápido ni sencillo.

7.1. Técnicas de muestreo

Una muestra debe ser un subconjunto de la población, representativo de la misma. Vamos a comentar algunas de las técnicas más habituales de muestreo.

Muestreo con reemplazamiento es el que se realiza cuando un elemento tomado de la población vuelve de nuevo a ella para poder volver a ser elegido. En esta situación, cada miembro de la población puede seleccionarse más de una vez.

Muestreo sin reemplazamiento es el que se realiza sin devolver a la población los elementos que se van eligiendo para construir la muestra. En esta situación, cada

miembro de la población solo puede seleccionarse una vez.

Muestreo no aleatorio es el que se realiza de modo que no todos los elementos de la población tienen la misma probabilidad de ser elegidos.

Con este tipo de muestreo la representatividad de la muestra es escasa y las inferencias poco válidas.

Dentro de este tipo de muestreo se encuentran los muestreos: *opinático* (elección subjetiva por considerar el elemento representativo); *por cuotas* (se obliga a elegir un cierto número de elementos con una característica determinada); *semialeatorio* (en alguna fase del muestreo aleatorio, se permite al entrevistador la elección del elemento que formará parte de la muestra) y *por rutas* (se suele utilizar en encuestas de opinión y consiste en indicar pautas para la elección del itinerario a seguir y que llevará al individuo encuestado).

Muestreo aleatorio es el que se realiza teniendo en cuenta que todos los individuos de la población tienen la misma probabilidad de ser elegidos en la muestra. Con este tipo de muestreo, las muestras son representativas, es posible conocer los errores cometidos y se pueden hacer inferencias.

El muestreo aleatorio es el que más nos interesa y será el que utilicemos siempre que podamos. Existen tres tipos de muestreo aleatorio:

Muestreo aleatorio simple Para elegir una muestra se parte de una lista con todos los elementos de la población y del mismo se seleccionan los n elementos que forman la muestra.

La elección de estos n elementos se pude hacer de varias formas:

- Asignando un número a cada elemento de la población y luego eligiendo al azar n números (ya sea metiéndolos en una urna y sacando n papeles, o generando una variable discreta equiprobable). Esto da lugar a un muestreo sin reemplazamiento.
- Mediante una tabla de números aleatorios. Tendremos que decidir si queremos un muestreo con reemplazamiento (un elemento puede estar más de una vez en la muestra), o no.

Este muestreo aleatorio es el más sencillo de todos y sirve de base para los otros dos.

Muestreo aleatorio sistemático es una variedad del muestreo aleatorio simple. Consiste en, conocido el tamaño de la población, N, y de la muestra, n, dividir N entre n, y el resultado del cociente, k, nos indica que debemos seleccionar los elementos de la muestra de k en k.

Este tipo de muestreo tiene la ventaja de que solo hay que elegir aleatoriamente el primer elemento de la muestra, pero tiene el problema de que si hay periodicidad en los datos, la muestra resultante puede que no sea representativa.

Muestreo aleatorio estratificado Se realiza dividiendo la población en subgrupos o estratos homogéneos y tomando, en cada uno de ellos, una muestra aleatoria simple.

El procedimiento utilizado para determinar el número de elementos que se toman en cada estrato se llama **afijación**. Los más habituales son:

- Afijación simple: se toma el mismo número de elementos en cada estrato.
- Afijación proporcional: el número de elementos es proporcional al tamaño del estrato dentro de la población.

  En general, si queremos tomar una muestra de tamaño n en una población de tamaño N, para el i-ésimo estrato, de tamaño N_i, tendremos que tomar una muestra de tamaño: $n_i = n\frac{N_i}{N}$.

Existen otros tipos de muestreo aleatorio aunque no vamos a verlos.

7.2. Tamaño de la muestra

7.2.1. Para la estimación de una media

Una cuestión importante a la hora de seleccionar una muestra es determinar el tamaño de la misma.

En muchas ocasiones esta tarea no es nada sencilla, e incluso lo único que podemos hacer es una estimación del tamaño mínimo que debe tener.

Empecemos por el principio, ¿por qué no nos vale con cualquier tamaño de muestra y en qué influye dicho tamaño?

Para empezar, debemos tener claro que se elige una muestra cuando tenemos una población tan grande que no podemos abarcarla, o bien lo suficientemente grande para que sea muy costoso el acceder a todos los elementos de la misma. En estos casos queremos elegir un subconjunto representativo (muestra aleatoria) y a la vez que no nos suponga un gasto excesivo.

Por otra parte, la muestra debe ser lo suficientemente grande como para que los resultados obtenidos a partir de ella sean fiables. Esta fiabilidad viene medida por el error máximo que estamos dispuestos a admitir, EM, y su probabilidad asociada (α).

Por lo tanto, lo que buscamos al elegir el tamaño muestral es **evitar un gasto excesivo y conseguir resultados fiables**.

El problema de determinar el tamaño de la muestra, como se ha dicho no es nada sencillo, pero, *en algunos casos*, podemos dar valores que nos garanticen una determinada fiabilidad de los resultados.

La idea es la siguiente:

Si la población es **Normal**, y la **varianza poblacional es conocida**, la distribución de la media muestral es:

$$\bar{\zeta} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

En este caso, el intervalo de confianza para la media poblacional, con un nivel de confianza $1-\alpha$, o lo que es lo mismo para un nivel de significación α, es:

$$IC(\mu) = \left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Es decir que: $IC(\mu) = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \bar{x} \pm EM$

EM es el **error muestral**.

Luego, el error muestral, que es el error máximo que se puede cometer, para un nivel de confianza $1-\alpha$, es:

$$EM = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Entonces:

Para calcular el intervalo de confianza para la media, o para realizar un contraste de hipótesis para la media, cuando:

- la población sigue una **distribución Normal** para la variable considerada,
- la **varianza poblacional es conocida**,
- estamos dispuestos a asumir una **probabilidad de error** α,
- y determinamos que el **error máximo que estamos dispuestos a aceptar es EM**,

entonces, el tamaño de la muestra debe ser:

$$n = \frac{z_{\alpha/2}^2\sigma^2}{EM^2}$$

En el caso de que la varianza poblacional sea desconocida (que es lo más habitual), sabemos que el intervalo de confianza para la media, con un nivel de confianza $1-\alpha$, es:

$$IC(\mu) = \left(\bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right)$$

Esto significa que el error muestral es:

$$EM = t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

pero en esta expresión tanto s como el valor de la t, dependen de n, por lo que es imposible calcularlos.

Lo que se hace en estas ocasiones es calcular una **estimación del tamaño de la muestra**, utilizando información previa fiable sobre el valor de la varianza poblacional.

Utilizaremos la misma expresión, pero ahora **no podemos asegurar el resultado**, aunque será aproximado.

Para calcular el intervalo de confianza para la media, o para realizar un contraste de hipótesis para la media cuando:

- la población sigue una **distribución Normal** para la variable considerada,
- aunque la **varianza poblacional es desconocida** *tenemos información previa fiable sobre la misma* (habitualmente se utiliza la cuasivarianza muestral s^2),
- estamos dispuestos a asumir una **probabilidad de error** α,
- y determinamos que el **error máximo que estamos dispuestos a aceptar es EM**,

entonces, el tamaño de la muestra debe ser:

$$n = \frac{z_{\alpha/2}^2 s^2}{EM^2}$$

En cualquier otro caso, el análisis es mucho más complicado y no lo vamos a ver.

7.2.2. Para la estimación de una proporción

En el caso de la **proporción poblacional**, el razonamiento es análogo:

Sabemos que cuando n es suficientemente grande ($n > 30$), el intervalo de confianza para la proporción poblacional, con un nivel de significación α, es:

$$IC(p) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Es decir que el error muestral es:

$$EM = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Lo que significa que el tamaño mínimo de la muestra debe ser:

$$n = z_{\alpha/2}^2 \frac{\hat{p}(1-\hat{p})}{EM^2}$$

Es decir:

Para calcular el intervalo de confianza para la proporción poblacional, o para realizar un contraste de hipótesis para la proporción poblacional, si estamos dispuestos a asumir una **probabilidad de error** α, y determinamos que el **error máximo** que estamos dispuestos a aceptar es **EM**, entonces, el tamaño de la muestra debe ser:

$$n = z^2_{\alpha/2} \frac{\hat{p}(1-\hat{p})}{EM^2}$$

Este resultado es cierto siempre que obtengamos $n > 30$**.**

En el resultado anterior, se utiliza una proporción muestral previa (o estimada) como estimación de la proporción poblacional para poder realizar los cálculos.

Si no queremos o no podemos utilizar la aproximación de la proporción, nos podemos poner en la peor situación posible y determinar el *intervalo de confianza más grande posible*, para un nivel de confianza $1-\alpha$ que es el que se obtiene cuando **p=1/2**.

Para esta situación sabemos que:

Cuando n es suficientemente grande ($n > 30$), el intervalo de confianza para la proporción poblacional, con un nivel de significación α, es:

$$IC(p) = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{1}{4n}}$$

Es decir que el error muestral es: $EM = z_{\alpha/2}\sqrt{\frac{1}{4n}}$

Lo que significa que el tamaño mínimo de la muestra debe ser:

$$n = z^2_{\alpha/2} \frac{1}{4EM^2}$$

Es decir que:

Si no queremos o no podemos utilizar la aproximación de la proporción, para calcular el intervalo de confianza para la proporción poblacional, o para realizar un contraste de hipótesis para la proporción poblacional, si estamos dispuestos a asumir una **probabilidad de error** α, y determinamos que el **error máximo** que estamos dispuestos a aceptar es **EM**, entonces, el tamaño de la muestra debe ser:

$$n = z^2_{\alpha/2} \frac{1}{4EM^2}$$

Este resultado es cierto siempre que obtengamos $n > 30$**.**

7.2.3. Para la estimación de una diferencia de medias

En el caso de la diferencia de medias, solo vamos a considerar el caso más elemental:

Sabemos que si tenemos muestras tomadas en Poblaciones Normales o poblaciones cualesquiera pero con muestras grandes y varianzas poblacionales conocidas, la distribución en el muestreo de la diferencia de medias cumple que:

$$\frac{\bar{\zeta}_1 - \bar{\zeta}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

Entonces, el intervalo de confianza para la diferencia de medias será:

$$IC(\mu_1 - \mu_2) = \bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Si tomamos **muestras del mismo tamaño** en las dos poblaciones: **n=m**, entonces:

$EM = z_{\alpha/2}\sqrt{\dfrac{\sigma_1^2 + \sigma_2^2}{n}}$ y en este caso: $n = z_{\alpha/2}^2\left(\dfrac{\sigma_1^2 + \sigma_2^2}{EM^2}\right)$

Es decir

Si trabajamos en Poblaciones Normales con varianzas conocidas y si tomamos **muestras del mismo tamaño** en ambas poblaciones, entonces, el tamaño muestral necesario en cada población, para que el error muestral de la diferencia de medias, con un nivel de confianza prefijado $1 - \alpha$, sea igual a una cantidad prefijada, EM, es:

$$n = z_{\alpha/2}^2\left(\frac{\sigma_1^2 + \sigma_2^2}{EM^2}\right)$$

7.2.4. Para la estimación de una diferencia de proporciones

Análogamente, para la diferencia de proporciones

Si tomamos **muestras del mismo tamaño** en ambas poblaciones, entonces, el tamaño muestral necesario para que el error muestral de la diferencia de proporciones, con un nivel de confianza prefijado $1 - \alpha$, sea igual a una cantidad prefijada, EM, es:

$$n = z_{\alpha/2}^2\left(\frac{p_1(1-p_1) + p_2(1-p_2)}{EM^2}\right)$$

Igual que en el caso de una proporción, para poder aplicar esta fórmula, tendremos que utilizar alguna estimación «fiable» de las proporciones poblacionales ya que estas son desconocidas.

O bien, en el caso más extremo ($p_1 = p_2 = 1/2$):

Si tomamos **muestras del mismo tamaño** en ambas poblaciones, entonces, el tamaño muestral necesario para que el error muestral de la diferencia de proporciones, con un nivel de confianza prefijado $1-\alpha$, sea igual a una cantidad prefijada, EM, es:

$$n = z_{\alpha/2}^2 \left(\frac{1}{2EM^2} \right)$$

Tema 8

Estadística Descriptiva bidimensional

Hasta ahora, prácticamente todo lo que hemos visto se refería al estudio de una única variable estadística o aleatoria. Vamos a ver aquí cómo abordar el estudio de una variable bidimensional como una extensión de lo ya visto en el caso unidimensional.

La mayoría de las veces, al estudiar una población, se estudian dos o más características simultáneamente. Cada observación dará lugar por tanto a dos o más números (suponiendo que las características son cuantitativas). La variable estadística correspondiente se denomina: **variable bidimensional o multidimensional**.

Nosotros nos vamos a limitar al estudio de dos características, por lo que nos centraremos en las variables bidimensionales.

El análisis de las distribuciones de dos o más dimensiones tiene por objetivo general el estudio de la existencia o no de algún tipo de asociación, dependencia o covariación entre las distintas variables.

8.1. Tablas de frecuencias

Igual que hacíamos en el caso unidimensional, una vez que hemos recogido nuestra masa de datos, el primer paso será intentar resumir esta información, para lo cual construiremos una tabla de frecuencias.

Si tenemos una variable X, con valores: $x_1, x_2, .., x_k$, y otra variable Y con valores $y_1, y_2, .., y_m$, para cada elemento de la población tendremos una observación bidimensional (x_i, y_j).

Llamaremos:

Frecuencia absoluta conjunta bidimensional al número de veces que se presenta conjuntamente el par de valores (x_i, y_j), y se representa por n_{ij}.

Frecuencia relativa conjunta bidimensional a la proporción de veces que se presenta el par (x_i, y_j) y se calcula como el cociente entre la frecuencia absoluta bidimensional y el número total de datos: $f_{ij} = \dfrac{n_{ij}}{N}$

Distribución bidimensional al conjunto formado por los pares de valores de los caracteres (x_i, y_j), asociado a sus frecuencias absolutas: (x_i, y_j, n_{ij}), o a las relativas.

Podemos construir una tabla de frecuencias con las variables X e Y y su frecuencia conjunta, es decir, para cada par de valores, su frecuencia: (x_i, y_j, n_{ij}).

Otra forma de disponer los datos es la conocida como **tabla de doble entrada** (si es de caracteres cualitativos o atributos se denomina **tabla de contingencia**).

Igual que en el caso de variables unidimensionales, podemos distinguir entre distribuciones agrupadas en intervalos o no agrupadas.

Tabla de doble entrada genérica:

NO AGRUPADA:

Y / X	y_1	y_2	y_3	...	y_m
x_1	n_{11}	n_{12}	n_{13}	...	n_{1m}
x_2	n_{21}	n_{22}	n_{23}	...	n_{2m}
...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	n_{k3}	...	n_{km}

AGRUPADA:

Y / X	$(l'_0, l'_1]$	$(l'_1, l'_2]$	$(l'_2, l'_3]$	...	$(l'_{m-1}, l'_m]$
$(l_0, l_1]$	n_{11}	n_{12}	n_{13}	...	n_{1m}
$(l_1, l_2]$	n_{21}	n_{22}	n_{23}	...	n_{2m}
...	...	...	...	...	...
$(l_{k-1}, l_k]$	n_{k1}	n_{k2}	n_{k3}	...	n_{km}

Ejemplo: Sea una población de 96 familias, para la que se han medido las siguientes variables:

X= número de personas activas en la familia.

Y= tamaño de la familia (número de miembros).

Y / X	1	2	3	4	5	6	7	8	
1	7	10	11	16	8	1	1	0	(54)
2	0	2	5	6	6	2	0	0	(21)
3	0	0	1	6	4	3	1	1	(16)
4	0	0	0	0	2	1	1	1	(5)
	(7)	(12)	(17)	(28)	(20)	(7)	(3)	(2)	96

Extraemos información de la tabla:

- $N = \sum_{i,j} n_{ij}$: la suma de todas las frecuencias, coincide con el número total de observaciones.

- $n_{43} = 0$; El número de familias de 3 miembros con 4 personas activas es 0.

- $n_{25} = 6$; hay 6 familias de 5 miembros en las que 2 están en activo.

- $f_{25} = \frac{6}{96} = 0.0625$; de las 96 familias, hay 6 familias de 5 miembros con 2 en activo. O bien, 0.0625 es la proporción de familias de 5 miembros con 2 en activo, en el total de las 96 familias. Multiplicando por 100 se obtiene el porcentaje (6.25 %).

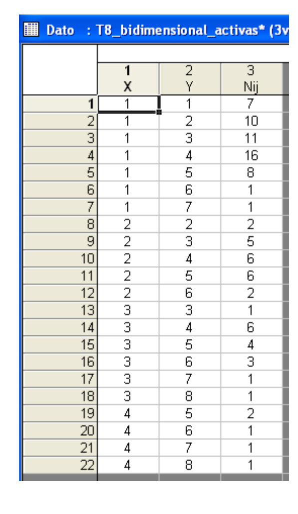
8.2. Gráficos

Las distribuciones bidimensionales se pueden representar gráficamente en el espacio de tres dimensiones.

En este caso en el eje vertical se representan las frecuencias y en el plano horizontal los valores de las variables X e Y.

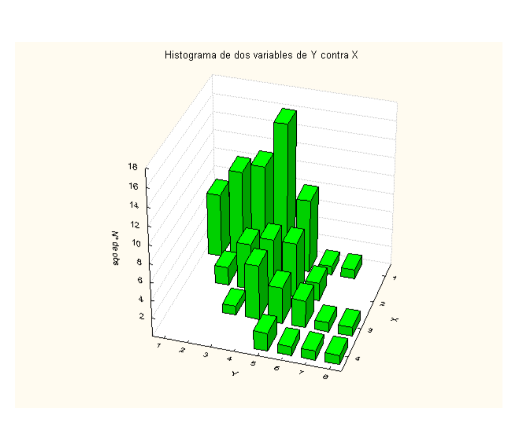
Por ejemplo, la representación gráfica de una distribución genérica puede ser:

Gráfico de barras de la variable bidimensional, o **Histograma de datos categóricos** en variables no agrupadas (representan lo mismo):

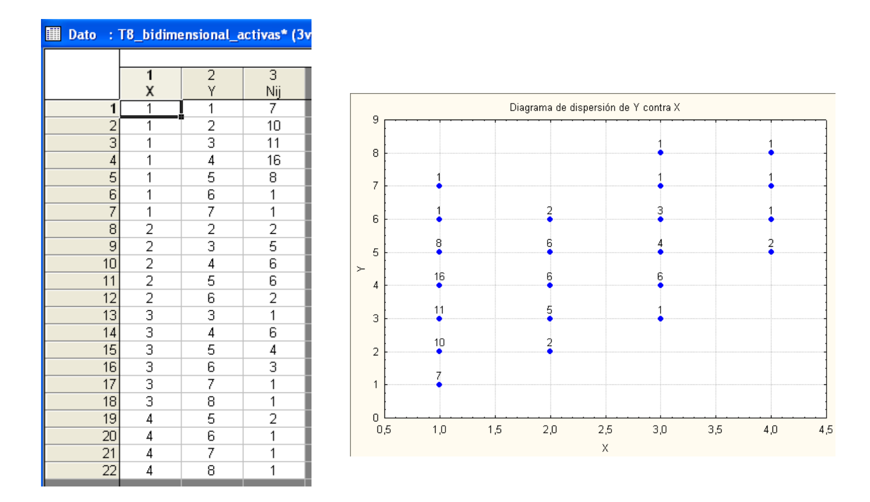


Dato : T8_bidimensional_activas* (3v

	1 X	2 Y	3 Nij
1	1	1	7
2	1	2	10
3	1	3	11
4	1	4	16
5	1	5	8
6	1	6	1
7	1	7	1
8	2	2	2
9	2	3	5
10	2	4	6
11	2	5	6
12	2	6	2
13	3	3	1
14	3	4	6
15	3	5	4
16	3	6	3
17	3	7	1
18	3	8	1
19	4	5	2
20	4	6	1
21	4	7	1
22	4	8	1

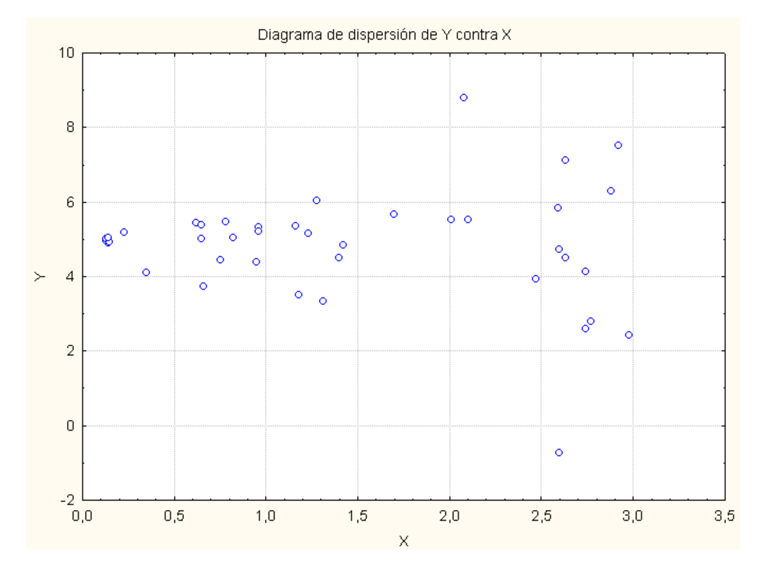


Nube de puntos o **Diagrama de Dispersión**:

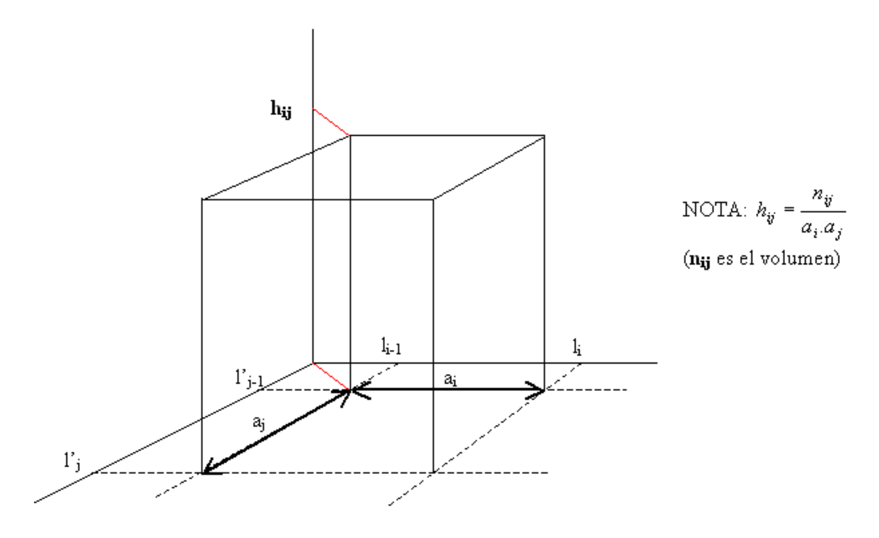

Dato : T8_bidimensional_activas* (3v

	1 X	2 Y	3 Nij
1	1	1	7
2	1	2	10
3	1	3	11
4	1	4	16
5	1	5	8
6	1	6	1
7	1	7	1
8	2	2	2
9	2	3	5
10	2	4	6
11	2	5	6
12	2	6	2
13	3	3	1
14	3	4	6
15	3	5	4
16	3	6	3
17	3	7	1
18	3	8	1
19	4	5	2
20	4	6	1
21	4	7	1
22	4	8	1

El Diagrama de dispersión es más práctico cuando todos los pares de valores tienen frecuencia 1:



En el caso de variables agrupadas en intervalos, podemos dibujar el Histograma en el espacio, aunque puede ser complicado.



8.3. Distribuciones marginales y condicionadas

8.3.1. Distribuciones marginales

En una distribución bidimensional, (X, Y), se pueden considerar las distribuciones de cada una de las variables componentes. A estas distribuciones se les llama **distribuciones marginales** y vienen definidas por los valores que toma una variable y la frecuencia de los mismos, al margen (de ahí el nombre) de los que tome la otra.

Consideremos, por ejemplo:

$X \backslash Y$	y_1	y_2	y_3
x_1	n_{11}	n_{12}	n_{13}
x_2	n_{21}	n_{22}	n_{23}
x_3	n_{31}	n_{32}	n_{33}
x_4	n_{41}	n_{42}	n_{43}

Para saber el número de veces que ha aparecido el valor x_1, basta con sumar las frecuencias correspondientes a dicha fila. Llamando a ese valor $n_{1\bullet}$, podemos poner:

$n_{1\bullet} = n_{11} + n_{12} + n_{13} = \sum_{j=1}^{3} n_{1j}$; en general: $n_{i\bullet} = \sum_{j=1}^{m} n_{ij}$

Análogamente, para la variable Y, tenemos:

$n_{\bullet j} = n_{1j} + n_{2j} + n_{3j} + n_{4j} = \sum_{i=1}^{4} n_{ij}$; y en general: $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$

Además:

$$\sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{m} n_{\bullet j} = \sum_{i} \sum_{j} n_{ij} = \sum_{j} \sum_{i} n_{ij} = N$$

Podemos completar la tabla de doble entrada (llamada también de correlación) con esta información:

$X \backslash Y$	y_1	y_2	y_3	$n_{i\bullet}$
x_1	n_{11}	n_{12}	n_{13}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	n_{23}	$n_{2\bullet}$
x_3	n_{31}	n_{32}	n_{33}	$n_{3\bullet}$
x_4	n_{41}	n_{42}	n_{43}	$n_{4\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	N

Por ejemplo, para las variables: X=estado civil e Y=tipo de residencia

Y \ X	Urbano	Intermedio	Rural	$n_{i\bullet}$
Soltero	191	180	129	500
Casado	211	196	131	538
Viudo	40	28	35	103
Otros	8	2	0	10
$n_{\bullet j}$	450	406	295	1151

8.3.2. Distribuciones condicionadas

En ocasiones estaremos interesados en analizar un cierto subconjunto de la población total, es decir, solo aquellos elementos que cumplen una determinada condición.

Así en la distribución conjunta de X e Y, podemos estar interesados solo en el análisis de la variable X, pero refiriéndonos únicamente a aquellos elementos para los que la variable Y toma un determinado valor: y_r.

Por ejemplo, en el caso de las familias, nos puede interesar hacer un análisis, no de todas las familias (96), sino solo de aquellas en las que hay menos de 3 miembros activos, o solo de aquellas que tienen 5 miembros.

La variable X, sujeta a la condición de que la Y tome el valor concreto y_r , la representaremos por: $X|_{Y=y_r}$.

De igual forma se define la variable Y condicionada al hecho de que la variable X tome el valor x_s : $Y|_{X=x_s}$

La **frecuencia relativa condicionada** $f(x_i|y_j)$, se define como la frecuencia relativa con que se presenta x_i , dentro del subconjunto en el que $Y = y_j$.

Es decir:

$$f(x_i|y_j) = \frac{f(x_i, y_j)}{f(y_j)} = \frac{f_{ij}}{f_{\bullet j}} = \frac{n(x_i, y_j)}{n(y_j)} = \frac{n_{ij}}{n_{\bullet j}}$$

Análogamente

$$f(y_j|x_i) = \frac{f(x_i, y_j)}{f(x_i)} = \frac{f_{ij}}{f_{i\bullet}} = \frac{n(x_i, y_j)}{n(x_i)} = \frac{n_{ij}}{n_{i\bullet}}$$

Nuestro universo son los valores que cumplen la condición, por lo que la suma de todas las frecuencias condicionadas de X para un valor dado de Y es igual a 1.

$$\sum_{i=1}^{k} f(x_i|y_j) = \frac{\sum_{i=1}^{k} f_{ij}}{f_{\bullet j}} = \frac{f_{\bullet j}}{f_{\bullet j}} = 1$$

La distribución condicionada de la variable Y, dado un valor $X = x_i$, se define por los valores que toma la variable Y, y las frecuencias condicionadas de Y asociadas a dichos valores. Análogamente se define una distribución condicionada de la variable X.

Las distribuciones condicionadas son, en realidad, unas distribuciones unidimensionales en las que se pueden calcular las mismas características que en estas últimas.

Así, en la distribución condicionada de Y, dado $X = x_i$, la media vendrá dada por la siguiente expresión:

$$\bar{y}|x_i = \sum_{j=1}^{m} y_j f(y_j|x_i)$$

Y la varianza vendrá dada por:

$$\mathrm{S}'^2_{y|x_i} = \sum_{j=1}^{m} (y_j - \bar{y}|x_i)^2 f(y_j|x_i)$$

En el ejemplo de las familias:

Y / X	1	2	3	4	5	6	7	8
1	7	10	11	16	8	1	1	0
2	0	2	5	6	6	2	0	0
3	0	0	1	6	4	3	1	1
4	0	0	0	0	2	1	1	1

Para las familias de 5 miembros, el número medio de personas activas es de 2 con una varianza de 1:

x_i	$n_i\|_{y=5}$	$x_i n_i\|_{y=5}$	$x_i^2 n_i\|_{y=5}$
1	8	8	8
2	6	12	24
3	4	12	36
4	2	8	32
suma	20	40	100

$$\bar{x}|_{y=5} = \frac{40}{20} = 2$$

$$a_2(x|_{y=5}) = \frac{100}{20} = 5$$

$$\mathrm{S}'^2_{x|_{y=5}} = 5 - 2^2 = 1$$

Análogamente: Para las familias de 6 miembros o más, el número medio de personas activas es de $\frac{11}{4} = 2.75$ con una varianza de $\frac{49}{48}$:

x_i	$n_i\|_{y\geq 6}$	$x_i n_i\|_{y\geq 6}$	$x_i^2 n_i\|_{y\geq 6}$
1	2	2	2
2	2	4	8
3	5	15	45
4	3	12	48
suma	12	33	103

$$\bar{x}|_{y\geq 6} = \frac{33}{12} = \frac{11}{4} = 2.75$$

$$a_2(x|_{y\geq 6}) = \frac{103}{12}$$

$$S'^2_{x|_{y\geq 6}} = \frac{103}{12} - \left(\frac{11}{4}\right)^2 = \frac{49}{48}$$

Ejemplo:

Sea la siguiente distribución:

		60 50-70	**80** 70-90	**100** 90-110	
	X \ Y				$n_{i\bullet}$
160	150-170	35	30	5	70
180	170-190	3	48	29	80
	$n_{\bullet j}$	38	78	34	150

1. Frecuencias conjuntas: $n_{23} = 29$; $n_{12} = 30$

2. Frecuencias relativas:

En tanto por uno	En tanto por ciento
$f_{11} = \frac{35}{150} = 0.2\hat{3}$	$f_{11} = 23.\hat{3}\,\%$
$f_{22} = \frac{48}{150} = 0.32\hat{3}$	$f_{22} = 32.\hat{3}\,\%$

3. Frecuencias marginales: $n_{\bullet 2} = 78$; $n_{2\bullet} = 80$

4. Frecuencias relativas marginales:

En tanto por uno	En tanto por ciento
$f_{\bullet 3} = \frac{34}{150} = 0.22\hat{6}$	$f_{\bullet 3} = 22.\hat{6}\,\%$
$f_{1\bullet} = \frac{70}{150} = 0.4\hat{6}$	$f_{1\bullet} = 46.\hat{6}\,\%$

5. Frecuencias relativas condicionadas:

En tanto por uno	En tanto por ciento
$f(x = 160\ \|_{y=80}) = \frac{30}{78} = 0.3846$	$38.46\,\%$
$f(y = 100\ \|_{x=180}) = \frac{29}{80} = 0.3625$	$36.25\,\%$
$f(x = 180\ \|_{y<100}) = \frac{51}{116} = 0.4397$	$43.97\,\%$

6. Medias marginales y condicionadas:

$\bar{x} = \frac{160\times 70+180\times 80}{150} = 170.\hat{6}$

$\bar{x}\ |_{y=60} = \frac{160\times 35+180\times 3}{38} = 161.5789$

$\bar{y}\ |_{x=180} = \frac{60\times 3+80\times 48+100\times 29}{80} = 86.5$

8.4. La covarianza

¿Cómo sabemos si dos variables están relacionadas?

Para estudiar el grado de covariación o variación conjunta de dos variables calcularemos un coeficiente llamado COVARIANZA.

$$\mathrm{S}'_{XY} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{N}$$

Si realizamos cálculos en la expresión anterior, este coeficiente se puede obtener también como:

$$\mathrm{S}'_{XY} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij}}{N} - \bar{x}\bar{y}$$

(Esta expresión es más fácil de calcular)

El inconveniente de este coeficiente es que viene medido en el producto de las unidades de las dos variables.

Propiedades: como pasaba con la varianza, al ser un coeficiente calculado a través de las desviaciones respecto a la media, a este coeficiente **no le afectan los cambios de origen pero sí le afectan los cambios de escala**:

Si sabemos que X e Y son dos variables cuya covarianza es: $\mathrm{Cov}(X,Y) = \mathrm{S}'_{XY}$, y si tenemos dos variables U y V, que se obtienen como un cambio de origen y escala de las anteriores:

$u_i = a + bx_i$, y $v_j = c + dy_j$, entonces:

$$\mathrm{Cov}(U,V) = S'_{UV} = bdS'_{XY} = bd\mathrm{Cov}(X,Y)$$

Por último, diremos que dos variables X e Y están **incorreladas**, si su covarianza es cero: $\mathrm{S}'_{XY} = 0$.

Nota: en Inferencia (y los programas estadísticos) se usa la **Cuasicovarianza**, que se obtiene dividiendo por N-1 en lugar de N:

$$\mathrm{S}_{XY} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{N-1}$$

$$\mathrm{S}_{XY} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i y_j n_{ij}}{N-1} - \frac{N}{N-1}\bar{x}\bar{y}$$

8.5. Independencia

Intuitivamente, se puede afirmar que dos variables son independientes entre sí cuando los valores que toma una cualquiera de ellas no están afectados por los valores que toma la otra.

Definición: Dos variables son **independientes** si y solo si, la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales. Es decir:

$$\forall i, \forall j : f(x_i, y_j) = f(x_i)f(y_j)$$

Las siguientes proposiciones son equivalentes a la definición anterior:

- Diremos que la variable Y se distribuye independientemente de la variable X si y solo si, las frecuencias condicionadas de y_j (cualquiera que sea el valor de j) para los distintos valores de X, son iguales entre si. Es decir:

$$f(y_j|x_1) = f(y_j|x_2) = \cdots = f(y_j|x_k), \forall j$$

- Dos variables son independientes si y solo si:

$$f(y_j) = f(y_j|x_i), \forall i, \forall j$$

Definición: Dada una distribución bidimensional (X, Y), diremos que las variables son **dependientes** si y solo si, no son independientes. Es decir:

$$\exists i, j : f(x_i, y_j) \neq f(x_i)f(y_j)$$

Teorema: Independencia implica incorrelación, pero el recíproco no es cierto.

Es decir, que si dos variables son independientes, forzosamente su covarianza es cero (están incorreladas), pero puede ocurrir que la covarianza entre dos variables sea cero, y que **no sean** independientes.

Tema 9

Correlación y regresión lineal

En el tema anterior hemos examinado el concepto de dependencia entre dos variables, estableciendo el criterio para determinar si existe o no tal dependencia: $f_{ij} = f_{i\bullet} f_{\bullet j}$

Sin embargo, al investigador que está examinando la relación que existe entre dos variables no solo le interesa saber si las variables son independientes o no, sino que además, para profundizar en este análisis, será importante:

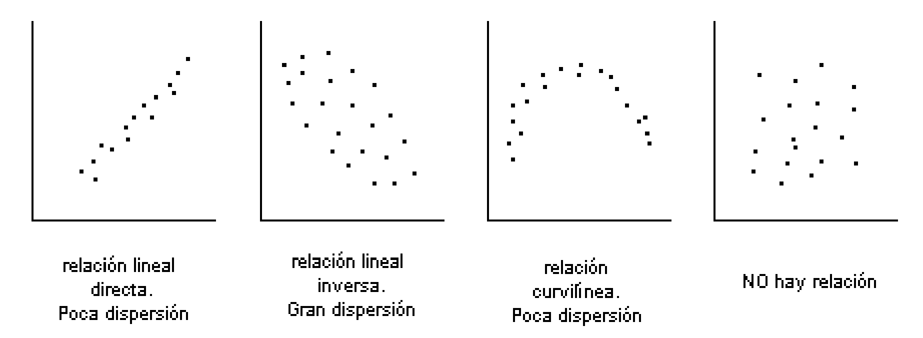
- medir el grado de asociación.
- conocer la forma concreta en la que se relacionan.

Precisamente para dar respuesta a estas dos cuestiones se han desarrollado las técnicas estadísticas de correlación y regresión.

La Correlación: estudia el grado de asociación entre las componentes de la variable estadística, y su objetivo es construir coeficientes que determinen si hay o no covariación.

La Regresión: se encarga de la determinación (si es posible) de aquella estructura de dependencia que mejor exprese el tipo de relación existente entre las componentes. Es decir, tratará de obtener (si es posible) una relación funcional entre las componentes: $y = f(x)$, en el caso bidimensional, o $y = f(x_1, x_2, \ldots, x_{k-1})$, en el caso k-dimensional.

Podemos encontrarnos con distintas situaciones:



Es importante señalar que la aplicación de estas técnicas estadísticas exige un análisis teórico previo de la relación que existe entre las variables objeto de estudio, pues lo contrario puede conducirnos a resultados absurdos.

Debe tenerse en cuenta que la **dependencia estadística** observada entre dos variables puede obedecer a tres motivos diferentes:

- **Al azar**.

  Podemos tomar dos variables para las que en principio no tiene ningún sentido estudiar su relación y descubrir que, casualmente, guardan una estrecha relación.

  Es famoso el ejemplo propuesto por G. M. Jenkins: el paralelo crecimiento del número de nacidos y el de cigüeñas en Baviera.

  Otro ejemplo: podemos encontrar que en los últimos 20 años, han crecido de forma paralela el número de divorcios en Suecia y los automóviles fabricados en España. Es evidente que carece de sentido llevar a cabo un estudio que relacione dichas variables.

- **Una tercera variable influye sobre las dos consideradas**.

  Por ejemplo, el aumento en el consumo de Whisky y la compra de automóviles, pueden moverse en una misma dirección a causa de la influencia que ejerce sobre los mismos la renta disponible.

  La relación entre la demanda de mobiliario y el aumento del precio del suelo, puede ser debida al aumento de la construcción.

- **Una variable influye en la otra** (la relación es de carácter causal).

  La relación que se establece entre las dos variables consideradas es de carácter causal. Por ejemplo: el gasto en ropa, realizado por una familia, viene influido por la renta que percibe.

Como parece lógico, las relaciones que nos interesará analizar son las de carácter causal, donde una variable llamada **explicativa** (o exógena), determina el comportamiento de otra variable llamada **explicada** (o endógena).

9.1. Correlación lineal

El grado de asociación existente entre dos variables, puede medirse mediante las técnicas de correlación. Estas técnicas nos proporcionan unos coeficientes que nos cuantifican ese grado de asociación.

El coeficiente de correlación más importante, es el

Coeficiente de correlación lineal de Pearson:

$$\mathrm{r}_{XY} = \frac{\mathrm{S}'_{XY}}{\mathrm{S}'_X \mathrm{S}'_Y}$$

El coeficiente de correlación lineal se utiliza para medir el **grado de asociación lineal** entre dos variables.

Importante: que este coeficiente sea cero solo significa que las variables no tienen ninguna relación lineal, pero pueden tener otro tipo de relación.

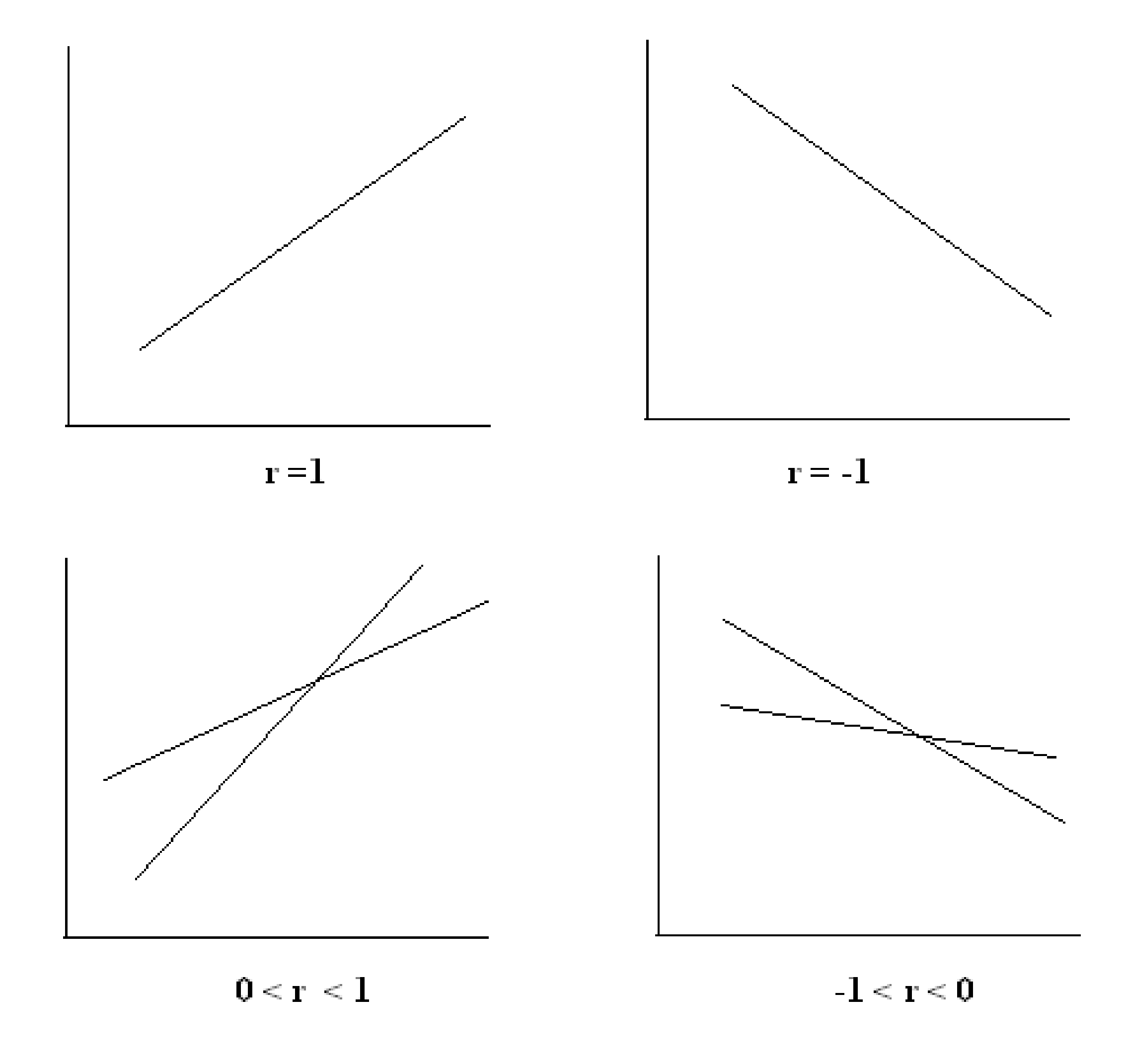
El valor de este coeficiente varía entre -1 y 1, y su signo dependerá del signo de la covarianza.

Cuanto más se aleje este coeficiente de cero (hacia el 1 o hacia el -1) mayor será la relación lineal entre las variables.

En el caso extremo de que el coeficiente sea 1 o -1, tendremos la máxima relación lineal, lo que significa que todos los puntos observados están alineados.

El signo del coeficiente, que es el signo de la covarianza, nos indica si la relación lineal entre las variables es directa (positivo: cuando los valores de una variable crecen, los de la otra también lo hacen) o inversa (negativo: cuando los valores de una variable crecen, los de la otra decrecen).

Gráficamente:



9.2. Regresión lineal

Una vez que hemos especificado que la relación entre dos variables es de tipo lineal, su formulación sería la siguiente: $y_i^* = a + bx_i$

El objeto de la regresión es la determinación del valor de los parámetros del modelo (en este caso a y b) a partir de un conjunto de observaciones sobre las variables.

La determinación de los parámetros se puede hacer por varios métodos; nosotros vamos a utilizar el **método de regresión minimocuadrática**.

La idea del método es la siguiente:

Supongamos que se ha determinado que existe una **relación lineal** entre las variables X=*renta familiar* e Y=*gasto en ropa*, entonces, el modelo teórico que define esta relación será:

$$y_i^* = a + bx_i$$

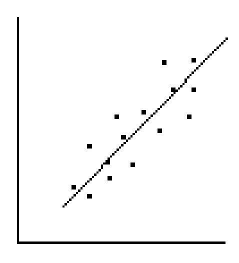
El gasto de una familia en ropa, puede estar influenciado especialmente por la renta, pero en ningún caso se puede esperar que esta variable explique completamente el gasto en el vestir. Existen otros factores: composición de la familia, clima, ideología, influencia de la moda,... que también ejercen una influencia en mayor o menor grado.

El número de estos factores puede ser infinito. Además, como la sencillez de los modelos es otra ventaja o propiedad a destacar, conviene incluir el mínimo número de variables posibles en los mismos. Para corregir esta anomalía expresaremos el modelo de la siguiente forma:

$$y_i = a + bx_i + e_i$$

donde e_i representa el error que cometemos al explicar el gasto en ropa en función únicamente de la renta, y que viene causado por múltiples efectos de procedencia muy dispar.

Por lo tanto, lo que tenemos es un conjunto de pares (x_i, y_i) a los que se quiere ajustar una recta:



$Y_i^* = a+bX_i$

Para cada valor de la variable X (renta), tenemos dos valores del gasto: un valor del gasto real (Y) y un valor del gasto teórico (Y^*). Por lo tanto, el error es la diferencia entre ambos:

$$e_i = y_i - y_i^*$$

Este error es el residuo o diferencia que queremos que sea lo menor posible.

Por lo tanto, **el problema consiste en encontrar los valores de los parámetros a y b del modelo, que minimicen el error**.

Pero ¿qué es lo que ocurre?, pues que como algunos errores son positivos y otros negativos se pueden compensar. Para evitar compensaciones, lo que haremos será: **Minimizar la suma de los cuadrados de los errores** (de ahí la denominación minimocuadrática):

$$\text{minimizar} \sum e_i^2$$

Para simplificar la notación, supondremos el caso de una distribución bidimensional de frecuencias unitarias, caso al que se pueden reducir los demás, sin más que repetir la pareja de valores tantas veces como nos indica su frecuencia.

Vamos a calcular los coeficientes de la recta que nos da los valores de la variable Y (variable explicada) en función de los valores de la variable X (variable explicativa), utilizando el método de regresión minimocuadrática.

Se trata de obtener los parámetros a y b de la recta (modelo teórico) $y_i^* = a + bx_i$, que mejor se ajusta a la nube de puntos, utilizando el método de ajuste de los mínimos cuadrados.

Es decir, que queremos **encontrar los valores de los parámetros a y b del modelo, de modo que minimizan la suma de los cuadrados de los errores**

$$\text{mín} \sum_{i=1}^{N} e_i^2 = \text{mín} \sum_{i=1}^{N} (y_i - y_i^*)^2 = \text{mín} \sum_{i=1}^{N} (y_i - a - bx_i)^2 = \text{mín}\, S(a,b)$$

El mínimo de esta función se obtiene cuando:

$b = \frac{\mathrm{S}'_{XY}}{\mathrm{S}'^2_X}$ y entonces $a = \bar{y} - b\bar{x}$

A la recta $Y^* = a + bX$ construida con estos parámetros se le denomina: **Recta de regresión de Y sobre X**, y se representa como: $Y|X$

La recta de regresión minimocuadrática, que explica los valores de la variable Y en función de los valores de la variable X se puede escribir de varias formas:

Cuando está escrita de la forma: $Y^* = a + bX$, diremos que está en **forma explícita** (es la forma habitual).

Sin embargo, cuando sustituimos los coeficientes de la recta por su valor:

$$y_i^* = \bar{y} - \frac{\mathrm{S}'_{XY}}{\mathrm{S}'^2_X}\bar{x} + \frac{\mathrm{S}'_{XY}}{\mathrm{S}'^2_X}x_i$$

obtenemos otra expresión de la misma recta:

$$y_i^* - \bar{y} = \frac{\mathrm{S}'_{XY}}{\mathrm{S}'^2_X}(x_i - \bar{x})$$

denominada **forma punto-pendiente**.

Análogamente, dada una variable bidimensional (X, Y), podemos construir la **recta de regresión minimocuadrática de X sobre Y**.

Es decir: el modelo teórico que explica los valores de la variable X en función de los valores de la variable Y.

Dicha recta será:

$$X|Y : X^* = a + bY, \text{ donde: } b = \frac{S'_{XY}}{S'^2_Y}, \text{ y } a = \bar{x} - b\bar{y}$$

Ejemplo:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
10	2	100	4	20
15	4	225	16	60
20	8	400	64	160
25	12	625	144	30
30	9	900	81	270
100	35	2250	309	810

Realizamos los cálculos:

$$\bar{x} = \frac{100}{5} = 20 \ ; \ \bar{y} = \frac{35}{5} = 7 \ ; \ S'^2_X = \frac{2250}{5} - 20^2 = 50$$

$$S'^2_Y = \frac{309}{5} - 7^2 = 12.8 \text{ y por último } S'_{XY} = \frac{810}{5} - 20 \times 7 = 22$$

Entonces las rectas de regresión son:

$$Y|X : y_i^* = \left(7 - \frac{22}{50} \times 20\right) + \frac{22}{50} x_i = -1.8 + 0.44 x_i$$

Es decir: $Y|X : y_i^* = -1.8 + 0.44x_i$, que explica el comportamiento de la variable Y en función de los valores que toma la variable X.

Y la otra recta de regresión es:

$$X|Y : x_i^* = \left(20 - \frac{22}{12.8} \times 7\right) + \frac{22}{12.8} y_i = 7.96875 + 1.71875 y_i$$

Es decir: $X|Y : x_i^* = 7.96875 + 1.71875y_i$, que explica el comportamiento de la variable X en función de los valores que toma la variable Y.

Propiedad de las rectas de regresión:

- Las rectas de regresión se cortan en el punto $(\bar{x}, \bar{y})$.

  Es decir, que este punto verifica la ecuación de las dos rectas.

Nota: como el signo de la pendiente de las rectas de regresión depende del signo de la covarianza, ambas rectas tiene la pendiente del mismo signo.

Es por esto que o bien ambas rectas son crecientes (pendientes positivas) o ambas rectas son decrecientes (pendientes negativas).

9.3. Análisis de la bondad del ajuste

Una vez realizado un ajuste, interesa constatar en qué medida queda explicada la variable endógena mediante el modelo estimado.

Un criterio bastante razonable para medir la bondad de cualquier ajuste es medir la proporción de varianza total explicada por el modelo.

Por ello, se toma como indicador de la bondad del ajuste, el cociente:

$$\mathrm{R}^2 = \frac{\mathrm{S}_{Y^*}'^2}{\mathrm{S}_Y'^2} = \textbf{Coeficiente de determinación}$$

El **coeficiente de determinación** R^2 , nos indica el tanto por uno de la variación de Y explicada por la variable X.

¿Cómo es este coeficiente en el caso de que ajustemos un modelo lineal?

En primer lugar, debemos notar que a partir de la distribución original (X, Y), podemos obtener otras dos distribuciones unidimensionales: las de y_i^* y la de e_i.

Cada una de estas nuevas distribuciones, tendrá sus características (su media y su varianza).

Además, en el caso de la regresión lineal minimocuadrática se cumple que:

$$\mathrm{S}_Y'^2 = \mathrm{S}_{Y^*}'^2 + \mathrm{S}_e'^2$$

Es decir: la Varianza total = Varianza explicada por la regresión + Varianza residual (no explicada).

Entonces, dividiendo la expresión anterior por $\mathrm{S}_Y'^2$, podemos expresar el coeficiente de determinación de la siguiente manera:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{S}_e'^2}{\mathrm{S}_Y'^2}$$

(este resultado es válido en el caso lineal, pero no con otros modelos)

Ejemplo: si $\mathrm{R}^2 = 0.99$, esto indica que el 99 % de la variación de Y está explicado por la variable X. Por lo tanto se trata de un buen ajuste.

Este resultado también nos indica, en el caso lineal, que el porcentaje de variación de la variable Y, que no está explicado por la variable X es del 1 %.

Los **límites de variación** se pueden ver fácilmente a través de la siguiente formulación:

$$\mathrm{R}^2 = 1 - \frac{\mathrm{S}_e^{\prime 2}}{\mathrm{S}_Y^{\prime 2}} = \frac{\mathrm{S}_{Y^*}^{\prime 2}}{\mathrm{S}_Y^{\prime 2}}$$

- Cuando la línea ajustada pasa por los puntos observados, todos los residuos serán nulos, y por lo tanto: $\mathrm{S}_e^{\prime 2} = 0$ y $\mathrm{R}^2 = 1$.

  Este será el máximo valor del coeficiente, y en consecuencia, la variación de Y viene totalmente explicada por X. Se trata de un ajuste perfecto.

- Por el contrario, si la varianza explicada por la regresión es nula, entonces: $\mathrm{R}^2 = 0$. Éste es el mínimo valor que puede tomar el coeficiente de determinación.

  El significado de este valor es que no existe ningún tipo de relación lineal entre las variables X e Y (que la variación de X no afecta para nada, **linealmente**, a la variación de Y)

  (No hay que olvidar que el hecho de que no haya relación lineal no significa que no pueda existir relación de algún otro tipo)

Cálculo de la varianza residual

Si tenemos el modelo lineal $y_i^* = a + bx_i$, como esto no es más que un cambio de origen y escala, sabemos que:

$$\mathrm{S}_{Y^*}^{\prime 2} = b^2 \mathrm{S}_X^{\prime 2} = \frac{\mathrm{S}_{XY}^{\prime 2}}{\mathrm{S}_X^{\prime 4}} \mathrm{S}_X^{\prime 2} = \frac{\mathrm{S}_{XY}^{\prime 2}}{\mathrm{S}_X^{\prime 2}} = b\mathrm{S}'_{XY}$$

Como en el caso lineal sabemos que se cumple que: $\mathrm{S}_Y^{\prime 2} = \mathrm{S}_{Y^*}^{\prime 2} + \mathrm{S}_e^{\prime 2}$

Entonces:

$$\mathrm{S}_e^{\prime 2} = \mathrm{S}_Y^{\prime 2} - \mathrm{S}_{Y^*}^{\prime 2} = \mathrm{S}_Y^{\prime 2} - \frac{\mathrm{S}_{XY}^{\prime 2}}{\mathrm{S}_X^{\prime 2}}$$

Esta fórmula nos permite calcular la varianza residual en función de las varianzas de las distribuciones marginales y de la covarianza de la distribución bidimensional.

9.4. Aplicaciones de la regresión

Son tres las aplicaciones más importantes de la regresión:

1. **LA PREDICCIÓN**. Esta es la aplicación más importante de la regresión.

   La predicción consiste en determinar, a partir del modelo estimado, el valor que toma la variable explicada, para un valor dado de la variable explicativa.

   Por ejemplo: Supongamos que en el caso de $Y =$ *gasto en ropa* y $X =$ *renta disponible*, tenemos que: $y_i^* = -20 + 0.044x_i$

   Entonces: si conocemos un valor concreto de la renta $x_0 = 3500$, podemos hacer una predicción teórica del gasto $y_0^* = -20 + 0.044 \times 3500 = 134$

   A partir de un modelo estimado, podemos hacer dos tipos de predicciones:

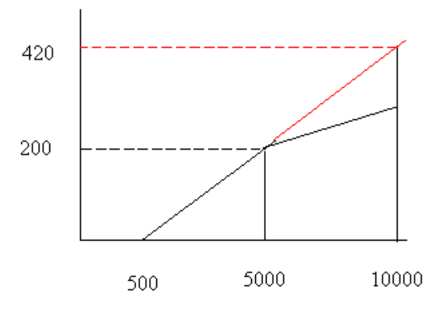
   - **Interpolación**: Para valores (x_0) de la variable explicativa que estén situados dentro del intervalo de variación de los datos.
   - **Extrapolación**: Para valores de la variable explicativa situados fuera del intervalo de variación de los datos.

   Al hacer predicciones a partir de un modelo dado, conviene hacer las siguiente observaciones:

   - La fiabilidad de los pronósticos para la variable endógena será tanto mejor cuanto mejor sea el ajuste, (es decir: cuanto mayor sea R^2, mejores serán los pronósticos), en el supuesto de que exista una relación causal entre las variables.
   - La fiabilidad de los valores pronosticados decrece a medida que el valor de la variable explicativa (en la que se basa la predicción) se aleja de la media ($\bar{x}$).

   Cuando hacemos predicciones para valores muy alejados del centro de gravedad de la distribución utilizada en el ajuste, se corre el peligro adicional de que no sea válido el modelo utilizado.

   Ejemplo: Puede ser aceptable el proponer una relación lineal entre el gasto en ropa y la renta disponible, para unos valores de la variable explicativa, digamos en $(500, 5000)$, pero a partir de ahí, es lógico pensar que el gasto en ropa tenga menor constante de proporcionalidad, es decir, que a partir de un valor en adelante, varía la pendiente (como se indica en la figura)



   Entonces, si efectuamos, a partir de nuestro modelo, la predicción para una renta de: $x_0 = 10000$, obtendríamos que y_0*=420, que no tendría ninguna fiablilidad.

Es decir, la **fiabilidad** *depende, tanto de la bondad del modelo como de la proximidad a los valores con los que ha sido estimado el mismo.*

2. **CALCULO DE FUNCIONES MARGINALES**.

   La función marginal representa la variación en unidades de la variable explicada al variar en una unidad la variable explicativa.

   El cálculo de la función marginal viene dado por: $marginal = \dfrac{dY}{dX}$

   En el caso del modelo lineal, $Y^* = a + bX$, la función marginal es el coeficiente angular (pendiente de la recta):

$$marginal = b = \frac{S'_{XY}}{S'^2_X}$$

3. **CALCULO DE ELASTICIDADES**.

   El coeficiente de elasticidad de Y con respecto a X será: la variación porcentual que experimenta Y al variar X en un 1 %.

   En Economía, se utiliza mucho el concepto de elasticidad de la demanda o de la oferta.

   Es un concepto ideado con objeto de indicar el grado en que la demanda (Q) o la oferta (O) responden a variaciones del precio del mercado (P). Depende esencialmente de las variaciones porcentuales, y es independiente de las unidades que utilicemos para medir Q (u O) y P.

   Ejemplo: La elasticidad-precio de la demanda, la definiríamos como la relación porcentual, o cambio porcentual en la cantidad demandada de un bien, que resulta del cambio en un 1 % en el precio.

   Analíticamente:

$$\varepsilon_{Y^*/X} = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{X}{Y}\frac{dY}{dX}$$

   En el caso del modelo lineal, $Y^* = a + bX$, la función de elasticidad es:

$$\varepsilon_{Y^*/X} = \frac{X}{Y}b$$

   Ejemplo: En el caso de Y=*gasto en ropa* y X=*renta disponible*, para el que hemos considerado que $y_i^* = -20 + 0.044x_i$, la función de elasticidad es:

$$\varepsilon_{Y^*/X} = 0.044\frac{X}{Y}$$

   Entonces, si calculamos la elasticidad para un valor concreto, por ejemplo para

   $x_0 = 600$, se tiene, según el modelo, que $y_0 = 6.4$

   por lo tanto, la elasticidad en este punto es: $\varepsilon_{Y^*/X} = 4.125$

   Esto significa que, si el valor $x_0 = 600$ aumenta en un 1 %, entonces el correspondiente valor de y_0 que habíamos obtenido, aumentará en un 4.125 %.

9.5. Ejemplo resuelto

Se quiere hacer un estudio sobre la relación entre la velocidad de los vehículos y el consumo de combustible.

En la siguiente tabla se muestran los consumos observados, en litros por cada 100 km, para cierto tipo de vehículos a distintas velocidades:

Velocidad (V)	80	80	80	80	120	120	120	120	140	140
Consumo (C)	4.5	5.8	5.0	5.5	6.0	6.6	7.2	6.5	7	8.5

1. ¿Se puede aceptar o no que existe una relación lineal entre el consumo y la velocidad?
2. Determina la ecuación lineal que nos da el consumo en función de la velocidad del vehículo.
3. ¿Qué porcentaje de la variabilidad del consumo no está explicado por la velocidad?
4. Para una velocidad de 110 km/h ¿cuál será el consumo estimado?, ¿es fiable esta estimación?
5. Si la velocidad anterior, 110 km/h se incrementa en un 1 %, ¿en qué porcentaje variará el consumo?

Solución:

1. ¿Se puede aceptar o no que existe una relación lineal entre el consumo y la velocidad?

   Para determinar si existe o no relación lineal entre las variables, calculamos el coeficiente de correlación lineal.

$$\mathrm{r}_{CV} = \frac{\mathrm{S'}_{CV}}{\mathrm{S'}_{C}\mathrm{S'}_{V}}$$

   Para obtenerlo debemos calcular las varianzas y la covarianza. Hacemos los cálculos previos:

V	C	V^2	C^2	CV
80	4.5	6400	20.25	360
80	5.8	6400	33.64	464
80	5.0	6400	25.00	400
80	5.5	6400	30.25	440
120	6.0	14400	36.00	720
120	6.6	14400	43.56	792
120	7.2	14400	51.84	864
120	6.5	14400	42.25	780
140	7.0	19600	49.00	980
140	8.5	19600	72.25	1190
1080	62.6	122400	404.04	6990

   Entonces:

$$\text{S'}_{CV} = \frac{6990}{10} - \frac{1080}{10}\frac{62.6}{10} = 22.92$$
$$\text{S}'^2_V = \frac{122400}{10} - \left(\frac{1080}{10}\right)^2 = 576$$
$$\text{S}'^2_C = \frac{404.04}{10} - \left(\frac{62.6}{10}\right)^2 = 1.2164$$

Luego:

$$\text{r}_{CV} = \frac{\text{S'}_{CV}}{\text{S'}_C\text{S'}_V} = \frac{22.92}{\sqrt{576 \times 1.2164}} = 0.86589$$

Esto significa que hay bastante relación lineal (el coeficiente está próximo a 1) y además la relación es directa (el coeficiente es positivo porque la covarianza es positiva), lo que indica que a medida que aumenta la velocidad aumenta también el consumo.

2. Determina la ecuación lineal que nos da el consumo en función de la velocidad del vehículo.

   Vamos a construir el modelo lineal que nos da el consumo en función de la velocidad: $C^* = a + bV$.

   Los coeficientes del modelo son:

   $b = \frac{\text{S'}_{CV}}{\text{S}'^2_V} = \frac{22.92}{576} = 0.03979$

   $a = \bar{c} - b\bar{v} = 6.26 - 0.03979 \times 108 = 1.96268$

   Entonces:

   $$C^* = 1.96268 + 0.03979V$$

3. ¿Qué porcentaje de la variabilidad del consumo no está explicado por la velocidad?

   Sabemos que el coeficiente de determinación representa la proporción de la variabilidad del consumo que está explicada por el modelo. Por lo tanto, la proporción no explicada será $1 - \text{R}^2$.

   $\text{R}^2_{CV} = \frac{\text{S}'^2_{CV}}{\text{S}'^2_C\text{S}'^2_V} = 0.86589^2 = 0.74977$

   El modelo explica un 74.977 % de la variabilidad del consumo en función de la velocidad, por lo tanto, como $1 - \text{R}^2 = 0.25023$:

   El porcentaje de la variabilidad del consumo que no está explicado por la velocidad es del 25.023 %.

4. Para una velocidad de 110 km/h ¿cuál será el consumo estimado?, ¿Es fiable esta estimación?

   Para hacer la estimación, basta utilizar la recta que acabamos de calcular:

   Si $V = 110$, entonces: $C^* = 1.96268 + 0.03979 \times 110 = 6.33958$

   Es decir, que a 110 km/h el consumo estimado es de unos 6.34 litros.

Para ver si esta estimación es fiable se deben comprobar dos cosas, por un lado la bondad del modelo y por otro la proximidad a los datos utilizados para la construcción del modelo. En nuestro caso:

- Bondad del ajuste: $R^2 = 0.74977$, el modelo es bastante bueno.
- Proximidad: 110 es un valor que está dentro del rango de velocidades utilizadas (de 80 a 140), es decir que estamos haciendo una interpolación y por lo tanto se cumple la condición de proximidad.

Entonces: **La predicción SÍ es fiable**

5. Si la velocidad anterior, 110 km/h se incrementa en un 1 %, ¿en qué porcentaje variará el consumo?

   Para responder a esta pregunta basta con calcular la elasticidad:

   $\varepsilon_{C/V} = \frac{dC}{dV}\frac{V}{C} = b\frac{V}{C} = 0.03979\frac{110}{6.33958} = 0.69041$

   El consumo aumenta en un 0.69 %.

   También se puede responder a esta pregunta directamente:

   Si $V = 110$, entonces: $C^* = 1.96268 + 0.03979 \times 110 = 6.33958$

   Si $V' = 110 \times 0.01 = 111.1$, entonces: $C^* = 1.96268 + 0.03979 \times 111.1 = 6.38335$

   La variación porcentual será:

   $$\varepsilon = \frac{6.38335 - 6.33958}{6.33958} \times 100 = 0.69042$$

   Es decir que **el consumo aumenta en un 0.69 %.**

Tema 10

Análisis estadístico de datos cualitativos

Este curso se ha centrado principalmente en el análisis de variables cuantitativas unidimensionales y bidimensionales.

¿Qué ocurre con las variables cualitativas? Con lo visto hasta ahora, prácticamente no podríamos pasar de construir una tabla de frecuencias y hacer alguna interpretación frecuentista de la misma o realizar algún gráfico. La parte de la Estadística que nos permite analizar las cualidades o características no medibles es la Estadística de atributos.

Sabemos que cuando las observaciones solo se pueden expresar en una escala nominal, lo único que podemos hacer es construir la tabla de frecuencias (contar las apariciones de cada valor), y en este caso, la única medida que nos sirve de resumen es la moda. Sin embargo, cuando los valores de la variable corresponden a una escala ordinal, para resumir la información, además de la moda también podemos utilizar la mediana.

Nos podríamos plantear ahora, tal como hemos hecho en el tema anterior, si existe algún tipo de relación entre las variables para características de este tipo. En este tema vamos a ver, sin entrar en muchos detalles, **cómo podemos cuantificar, si existe, la relación entre dos variables cualitativas**.

Cuando los caracteres estudiados son susceptibles de ser ordenados de acuerdo con una determinada escala, podremos llegar a unos coeficientes de correlación que midan el grado de asociación entre las variables. Estos coeficientes están basados en los rangos u órdenes de las observaciones.

En el caso de observaciones nominales, estableceremos los llamados coeficientes de asociación y contingencia.

10.1. Correlación por rangos

Para dos **variables ordinales**, queremos medir su **grado de asociación**.

Lo haremos mediante el **coeficiente de correlación por rangos de Spearman**.

Cada variable tiene una serie de valores que pueden ser ordenados, por lo tanto, a cada uno de ellos le podemos asociar su correspondiente rango o número de *ranking* (en caso de empates utilizaremos el criterio del rango central).

El coeficiente que vamos a calcular se basa en la comparación de los rangos para las dos variables:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N}$$

donde d_i es la diferencia entre los rangos en las dos variables para cada caso.

Este coeficiente es muy fácil de calcular, aunque se emplea, sobre todo, cuando tenemos menos de 20 observaciones.

A este coeficiente también se le llama coeficiente de correlación ordinal.

Interpretación del valor ρ:

El valor de este coeficiente varía entre -1 y +1.

- Cuando la concordancia entre los rangos es perfecta, entonces las diferencias son todas nulas y por lo tanto el coeficiente es igual a 1.
- Cuando existe discordancia total, los pares de rangos vienen dados por: $(N;1),(N-1;2),\ldots,(1;N)$. En ese caso, $\displaystyle\sum_{i=1}^{N} d_i^2 = \frac{N^3-N}{3}$, por lo que el coeficiente valdrá -1.
- Cuando el coeficiente tiene valor cero, indica que no existe relación entre los rangos de ambas variables.

Ejemplo:

Los *ranking* de 5 hoteles, según su ocupación y su precio son los que se dan en la siguiente tabla. Calcula el coeficiente de correlación por rangos de Spearman, para medir el grado de asociación de ambas variables:

Hotel	*Ranking* en ocupación	*Ranking* en precio	Diferencia (d_i)	d_i^2
A	1	3	-2	4
B	2	4	-2	4
C	3	2	1	1
D	4	1	3	9
E	5	5	0	0
			Suma	18

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N^3 - N} = 1 - \frac{6 \times 18}{5^3 - 5} = 1 - \frac{108}{120} = 0.1$$

El coeficiente está próximo a cero, lo que nos indica que el *ranking* en ocupación tiene muy poco que ver con el *ranking* en precio.

10.2. Asociación entre caracteres nominales

La observación de dos caracteres nominales da lugar a una tabla de doble entrada, en la que n_{ij} indica el número de objetos o individuos que poseen simultáneamente las modalidades i-ésima del primer atributo y j-ésima del segundo.

A estas tablas se les denomina **tablas de contingencia**.

Las distribuciones de frecuencias de cada uno de los atributos también se denominan distribuciones marginales.

Una tabla de contingencia tiene la siguiente forma:

Atributo B / Atributo A		Modalidad				
		B_1	B_2	$\ldots$	B_k	Total
Modalidad	A_1	n_{11}	n_{12}	$\ldots$	n_{1k}	$n_{1\bullet}$
	A_2	n_{21}	n_{22}	$\ldots$	n_{2k}	$n_{2\bullet}$
	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
	A_h	n_{h1}	n_{h2}	$\ldots$	n_{hk}	$n_{h\bullet}$
Total		$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet k}$	N

10.2.1. Tablas de contingencia 2×2

La clasificación por atributos más sencilla es la dicotómica, es decir, aquella en la que cada atributo solo tiene dos modalidades posibles (mutuamente excluyentes).

Por ejemplo:

Tener trabajo (B) / Sexo (A)	B_1 (Sí)	B_2 (No)	Total
$A_1(mujer)$	n_{11}	n_{12}	$n_{1\bullet}$
$A_2(hombre)$	n_{21}	n_{22}	$n_{2\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	N

Se dice que dos atributos son **independientes** cuando entre ellos no existe ninguna influencia mutua.

En el caso del ejemplo, no hay influencia entre los dos atributos si la proporción de personas con trabajo entre las mujeres es igual a la proporción de personas con trabajo entre los hombres, e igual a la proporción de personas con trabajo al margen del sexo. Es decir:

$$\frac{n_{11}}{n_{1\bullet}} = \frac{n_{21}}{n_{2\bullet}} = \frac{n_{\bullet 1}}{N}$$

Haciendo operaciones se puede ver que esto es equivalente a decir que: $\begin{cases} n_{11} = \frac{n_{1\bullet}n_{\bullet 1}}{N} \\ n_{21} = \frac{n_{2\bullet}n_{\bullet 1}}{N} \end{cases}$

y lo mismo ocurre con las demás modalidades: $n_{ij} = \frac{n_{i\bullet}n_{\bullet j}}{N}, (i, j = 1, 2)$

Es decir, que si dos atributos son estadísticamente independientes, la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales respectivas:

$$f_{ij} = f_{i\bullet}f_{\bullet j}, (i, j = 1, 2)$$

Como concepto contrario a la independencia, tenemos el de **asociación o dependencia**.

Se dice que **dos atributos están asociados** cuando aparecen juntos en más (o en menos) ocasiones que las que cabría esperar si fuesen independientes.

En las tablas dicotómicas la asociación se suele medir entre las distintas modalidades de los atributos.

Para medir el grado de asociación entre dos modalidades de dos atributos existen distintos coeficientes. Nosotros vamos a utilizar el **coeficiente de asociación H**.

En la tabla 2x2, sabemos que dos modalidades A_1 y B_1 son independientes si

$n_{11} = \dfrac{n_{1\bullet}n_{\bullet 1}}{N}$.

Por lo tanto, se puede obtener una primera medida de asociación mediante la diferencia:

$$H = n_{11} - \frac{n_{1\bullet}n_{\bullet 1}}{N}$$

Si H=0 $\Rightarrow$ Los atributos son independientes.

Si $H > 0 \Rightarrow$ Existe una asociación positiva.

Si $H < 0 \Rightarrow$ Existe una asociación negativa.

Este coeficiente es muy sencillo, pero tiene el inconveniente de que su amplitud depende de los valores que tomen las frecuencias conjuntas, así que aunque sabremos si existe o no asociación, no podemos determinar si esta es grande o pequeña.

Ejemplo:

De 1000 estudiantes, 516 son hombres y el resto mujeres. De los primeros, 221 son fumadores, mientras que las mujeres fumadoras son 183. Construye la tabla de doble entrada correspondiente y determina si existe asociación o independencia entre los atributos *sexo* y *ser fumador*.

Para analizar la asociación entre «mujer» y «sí es fumador», utilizaremos el coeficiente de asociación H:

En primer lugar construimos la tabla:

Ser fumador (B) Sexo (A)	B_1 (Sí)	B_2 (No)	Total
$A_1(mujer)$	183	301	484
$A_2(hombre)$	221	295	516
Total	404	596	1000

$$H = n_{11} - \frac{n_{1\bullet}n_{\bullet 1}}{N} = 183 - \frac{484 \times 404}{1000} = -12.536$$

Esto significa, que según las observaciones de que se dispone, existe una asociación negativa entre las modalidades *ser mujer* y *ser fumadora.*

Por lo tanto, **sí que hay asociación** entre los atributos.

Nota: calcula el coeficiente en las demás situaciones, ¿qué ocurre?, ¿por qué?

Con el coeficiente anterior podemos determinar si existe o no asociación entre las variables pero no podemos cuantificarla.

Si necesitamos **cuantificar la relación** usaremos una medida de asociación llamada **Q de Yule**, que se calcula como sigue:

$$Q = \frac{n_{11}n_{22} - n_{21}n_{12}}{n_{11}n_{22} + n_{21}n_{12}}$$

y cuya interpretación es la siguiente:

- Si las variables son independientes: Q=0
- Si existe asociación positiva (entre A_1 y B_1, y por lo tanto entre A_2 y B_2): $Q > 0$.
- Si existe asociación negativa (entre A_1 y B_1, y por lo tanto entre A_2 y B_2): $Q < 0$.

Además, como este coeficiente varía entre -1 y 1, y alcanza estos valores extremos cuando hay una asociación perfecta, este coeficiente nos permite medir la intensidad y la dirección de la asociación.

En el ejemplo anterior:

$$Q = \frac{n_{11}n_{22} - n_{21}n_{12}}{n_{11}n_{22} + n_{21}n_{12}} = \frac{183 \times 295 - 221 \times 301}{183 \times 295 + 221 \times 301} = \frac{-12536}{120506} = -0.104$$

Esto nos indica que hay una **relación negativa, aunque muy pequeña**, entre *ser mujer* y *ser fumadora.*

10.2.2. Tablas de contingencia $h \times k$

Estas tablas se construyen cuando el primer atributo tiene **h** modalidades y el segundo atributo **k** modalidades.

Para analizar la **independencia** de los dos atributos, se utiliza el mismo criterio que en el caso anterior, de modo que los dos atributos son independientes si:

$$\forall i, j : n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{N}$$

En este caso, para **medir el grado de asociación** entre los atributos, se utilizan los siguientes coeficientes de contingencia:

Coeficiente de contingencia χ^2

Si denominamos n_{ij} a la frecuencia conjunta observada de las modalidades A_i del atributo A y B_j del atributo B y por E_{ij} a la frecuencia teórica que le correspondería si fuesen independientes $\left(\forall i, j : E_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{N}\right)$, definimos el coeficiente de contingencia χ^2 como:

$$\chi^2 = \sum_{i=1}^{h} \sum_{j=1}^{k} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

A este coeficiente se le denomina también **cuadrado de la contingencia**.

Este coeficiente es siempre positivo y si las variables fuesen independientes su valor sería cero.

El coeficiente de contingencia χ^2 se suele utilizar para **contrastar la hipótesis de independencia** entre los atributos.

En este sentido, se tiene en cuenta que el estadístico χ^2, sigue una distribución Ji cuadrado con $(h-1) \times (k-1)$ grados de libertad: $\chi^2_{(h-1)(k-1)}$.

(Al final del tema, se indica cómo manejar las tablas para esta distribución)

Por lo tanto, aceptaremos la independencia de las variables si el p-valor es mayor que el nivel de significación (α).

En este caso, el p-valor es: $p = P\{X > \chi^2, \text{ siendo } X \sim \chi^2_{(h-1)(k-1)}\}$

O lo que es equivalente, **se acepta la independencia si** el valor calculado (χ^2) verifica:

$$\chi^2 < \chi^2_{(h-1)(k-1),\alpha}$$

siendo α el nivel de significación.

El coeficiente anterior no es muy adecuado para constituir por sí mismo un coeficiente, dado que sus límites varían en función de los datos, por ello, Karl Pearson propuso el siguiente coeficiente:

Coeficiente de contingencia de K. Pearson:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Este coeficiente varía entre 0 y 1, de modo que:

- si los atributos son independientes el valor de C es 0.
- cuanto mayor sea el grado de asociación más se acerca a 1.

En realidad con el coeficiente de contingencia de Pearson, el valor máximo, 1, no se alcanza más que en el caso teórico de infinitas modalidades, pero este coeficiente nos permite cuantificar y comparar el grado de asociación.

10.3. La distribución Ji cuadrado

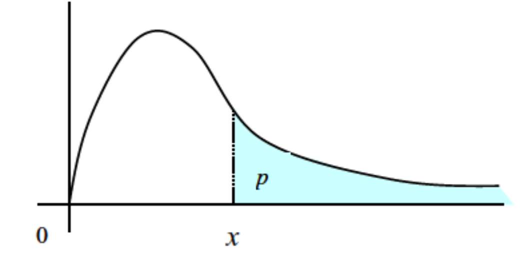
La distribución χ^2, Ji cuadrado de Pearson, es una distribución de probabilidad, continua y positiva y depende de un parámetro llamado grados de libertad.

Para indicar que una variable aleatoria X sigue una distribución Ji cuadrado con n grados de libertad, lo representaremos como: $X \sim \chi^2_n$

Como en otras distribuciones que hemos visto, los valores más usados están tabulados.

El manejo de la tabla de la Ji cuadrado es análogo al de la t de Student. Dados los grados de libertad y la probabilidad, la tabla nos indica el valor crítico, x, que deja a su derecha dicha probabilidad.

$P\{X > x\} = p =$ Área sombreada



Ejemplo:

a) Determina el valor crítico $\chi^2_{7,0.01}$ (valor crítico de una Ji cuadrado con 7 grados de libertad que deja a su derecha una probabilidad de 0.01).

b) Para una variable que se distribuye según una χ^2_3 ¿cuál es la probabilidad de que la variable tome un valor mayor que 0.584?

Buscamos en la tabla:

	p												
n	0.01	0.025	0.05	0.10	0.15	0.25	0.5	0.75	0.85	0.9	0.95	0.975	0.99
1	6.635	5.024	3.841	2.706	2.072	1.323	0.455	0.102	0.036	0.016	0.003932	0.000982	0.000157
2	9.210	7.378	5.991	4.605	3.794	2.773	1.386	0.575	0.325	0.211	0.103	0.051	0.020
3	11.345	9.348	7.815	6.251	5.317	4.108	2.366	1.213	0.798	0.584	0.352	0.216	0.115
4	13.277	11.143	9.488	7.779	6.745	5.385	3.357	1.923	1.366	1.064	0.711	0.484	0.297
5	15.086	12.833	11.070	9.236	8.115	6.626	4.351	2.675	1.994	1.610	1.145	0.831	0.554
6	16.812	14.449	12.592	10.645	9.446	7.841	5.348	3.455	2.661	2.204	1.635	1.237	0.872
7	18.475	16.013	14.067	12.017	10.748	9.037	6.346	4.255	3.358	2.833	2.167	1.690	1.239
8	20.090	17.535	15.507	13.362	12.027	10.219	7.344	5.071	4.078	3.490	2.733	2.180	1.646

RESPUESTAS: $\chi^2_{7,0.01} = 18.475$ y $P\{X > 0.584 | X \sim \chi^2_3\} = 0.9$

Tabla completa de la Ji cuadrado

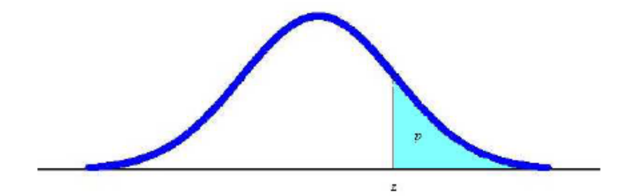
	p												
n	**0.01**	**0.025**	**0.05**	**0.10**	**0.15**	**0.25**	**0.5**	**0.75**	**0.85**	**0.9**	**0.95**	**0.975**	**0.99**
1	6.635	5.024	3.841	2.706	2.072	1.323	0.455	0.102	0.036	0.016	0.003932	0.000982	0.000157
2	9.210	7.378	5.991	4.605	3.794	2.773	1.386	0.575	0.325	0.211	0.103	0.051	0.020
3	11.345	9.348	7.815	6.251	5.317	4.108	2.366	1.213	0.798	0.584	0.352	0.216	0.115
4	13.277	11.143	9.488	7.779	6.745	5.385	3.357	1.923	1.366	1.064	0.711	0.484	0.297
5	15.086	12.833	11.070	9.236	8.115	6.626	4.351	2.675	1.994	1.610	1.145	0.831	0.554
6	16.812	14.449	12.592	10.645	9.446	7.841	5.348	3.455	2.661	2.204	1.635	1.237	0.872
7	18.475	16.013	14.067	12.017	10.748	9.037	6.346	4.255	3.358	2.833	2.167	1.690	1.239
8	20.090	17.535	15.507	13.362	12.027	10.219	7.344	5.071	4.078	3.490	2.733	2.180	1.646
9	21.666	19.023	16.919	14.684	13.288	11.389	8.343	5.899	4.817	4.168	3.325	2.700	2.088
10	23.209	20.483	18.307	15.987	14.534	12.549	9.342	6.737	5.570	4.865	3.940	3.247	2.558
11	24.725	21.920	19.675	17.275	15.767	13.701	10.341	7.584	6.336	5.578	4.575	3.816	3.053
12	26.217	23.337	21.026	18.549	16.989	14.845	11.340	8.438	7.114	6.304	5.226	4.404	3.571
13	27.688	24.736	22.362	19.812	18.202	15.984	12.340	9.299	7.901	7.042	5.892	5.009	4.107
14	29.141	26.119	23.685	21.064	19.406	17.117	13.339	10.165	8.696	7.790	6.571	5.629	4.660
15	30.578	27.488	24.996	22.307	20.603	18.245	14.339	11.037	9.499	8.547	7.261	6.262	5.229
16	32.000	28.845	26.296	23.542	21.793	19.369	15.338	11.912	10.309	9.312	7.962	6.908	5.812
17	33.409	30.191	27.587	24.769	22.977	20.489	16.338	12.792	11.125	10.085	8.672	7.564	6.408
18	34.805	31.526	28.869	25.989	24.155	21.605	17.338	13.675	11.946	10.865	9.390	8.231	7.015
19	36.191	32.852	30.144	27.204	25.329	22.718	18.338	14.562	12.773	11.651	10.117	8.907	7.633
20	37.566	34.170	31.410	28.412	26.498	23.828	19.337	15.452	13.604	12.443	10.851	9.591	8.260
21	38.932	35.479	32.671	29.615	27.662	24.935	20.337	16.344	14.439	13.240	11.591	10.283	8.897
22	40.289	36.781	33.924	30.813	28.822	26.039	21.337	17.240	15.279	14.041	12.338	10.982	9.542
23	41.638	38.076	35.172	32.007	29.979	27.141	22.337	18.137	16.122	14.848	13.091	11.689	10.196
24	42.980	39.364	36.415	33.196	31.132	28.241	23.337	19.037	16.969	15.659	13.848	12.401	10.856
25	44.314	40.646	37.652	34.382	32.282	29.339	24.337	19.939	17.818	16.473	14.611	13.120	11.524
26	45.642	41.923	38.885	35.563	33.429	30.435	25.336	20.843	18.671	17.292	15.379	13.844	12.198
27	46.963	43.195	40.113	36.741	34.574	31.528	26.336	21.749	19.527	18.114	16.151	14.573	12.879
28	48.278	44.461	41.337	37.916	35.715	32.620	27.336	22.657	20.386	18.939	16.928	15.308	13.565
29	49.588	45.722	42.557	39.087	36.854	33.711	28.336	23.567	21.247	19.768	17.708	16.047	14.256
30	50.892	46.979	43.773	40.256	37.990	34.800	29.336	24.478	22.110	20.599	18.493	16.791	14.953
31	52.191	48.232	44.985	41.422	39.124	35.887	30.336	25.390	22.976	21.434	19.281	17.539	15.655
32	53.486	49.480	46.194	42.585	40.256	36.973	31.336	26.304	23.844	22.271	20.072	18.291	16.362
33	54.776	50.725	47.400	43.745	41.386	38.058	32.336	27.219	24.714	23.110	20.867	19.047	17.074
34	56.061	51.966	48.602	44.903	42.514	39.141	33.336	28.136	25.586	23.952	21.664	19.806	17.789
35	57.342	53.203	49.802	46.059	43.640	40.223	34.336	29.054	26.460	24.797	22.465	20.569	18.509
36	58.619	54.437	50.998	47.212	44.764	41.304	35.336	29.973	27.336	25.643	23.269	21.336	19.233
37	59.893	55.668	52.192	48.363	45.886	42.383	36.336	30.893	28.214	26.492	24.075	22.106	19.960
38	61.162	56.896	53.384	49.513	47.007	43.462	37.335	31.815	29.093	27.343	24.884	22.878	20.691
39	62.428	58.120	54.572	50.660	48.126	44.539	38.335	32.737	29.974	28.196	25.695	23.654	21.426
40	63.691	59.342	55.758	51.805	49.244	45.616	39.335	33.660	30.856	29.051	26.509	24.433	22.164
41	64.950	60.561	56.942	52.949	50.360	46.692	40.335	34.585	31.740	29.907	27.326	25.215	22.906
42	66.206	61.777	58.124	54.090	51.475	47.766	41.335	35.510	32.626	30.765	28.144	25.999	23.650
43	67.459	62.990	59.304	55.230	52.588	48.840	42.335	36.436	33.512	31.625	28.965	26.785	24.398
44	68.710	64.201	60.481	56.369	53.700	49.913	43.335	37.363	34.400	32.487	29.787	27.575	25.148
45	69.957	65.410	61.656	57.505	54.810	50.985	44.335	38.291	35.290	33.350	30.612	28.366	25.901
50	76.154	71.420	67.505	63.167	60.346	56.334	49.335	42.942	39.754	37.689	34.764	32.357	29.707
55	82.292	77.380	73.311	68.796	65.855	61.665	54.335	47.610	44.245	42.060	38.958	36.398	33.570
60	88.379	83.298	79.082	74.397	71.341	66.981	59.335	52.294	48.759	46.459	43.188	40.482	37.485
65	94.422	89.177	84.821	79.973	76.807	72.285	64.335	56.990	53.293	50.883	47.450	44.603	41.444
70	100.425	95.023	90.531	85.527	82.255	77.577	69.334	61.698	57.844	55.329	51.739	48.758	45.442
75	106.393	100.839	96.217	91.061	87.688	82.858	74.334	66.417	62.412	59.795	56.054	52.942	49.475
80	112.329	106.629	101.879	96.578	93.106	88.130	79.334	71.145	66.994	64.278	60.391	57.153	53.540
85	118.236	112.393	107.522	102.079	98.511	93.394	84.334	75.881	71.589	68.777	64.749	61.389	57.634
90	124.116	118.136	113.145	107.565	103.904	98.650	89.334	80.625	76.195	73.291	69.126	65.647	61.754
95	129.973	123.858	118.752	113.038	109.286	103.899	94.334	85.376	80.813	77.818	73.520	69.925	65.898
100	135.807	129.561	124.342	118.498	114.659	109.141	99.334	90.133	85.441	82.358	77.929	74.222	70.065

Apéndice A

Tablas

Tabla de la Normal tipificada

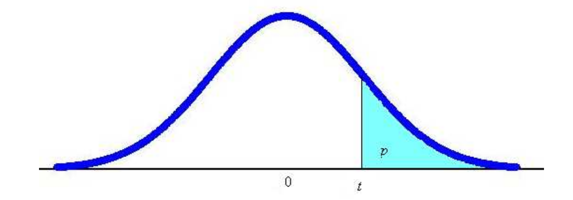
$P\{Z > z | Z \sim N(0,1)\}$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014

z	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
3	0.00135	$0.0^{3}968$	$0.0^{3}687$	$0.0^{3}483$	$0.0^{3}337$	$0.0^{3}233$	$0.0^{3}159$	$0.0^{3}108$	$0.0^{4}723$	$0.0^{4}481$
4	$0.0^{4}317$	$0.0^{4}207$	$0.0^{4}133$	$0.0^{5}854$	$0.0^{5}541$	$0.0^{5}340$	$0.0^{5}211$	$0.0^{5}130$	$0.0^{6}793$	$0.0^{6}479$
5	$0.0^{6}287$	$0.0^{6}170$	$0.0^{7}996$	$0.0^{7}579$	$0.0^{7}333$	$0.0^{7}190$	$0.0^{7}107$	$0.0^{8}599$	$0.0^{8}332$	$0.0^{8}182$
6	$0.0^{9}987$	$0.0^{9}530$	$0.0^{9}282$	$0.0^{9}149$	$0.0^{10}777$	$0.0^{10}402$	$0.0^{10}206$	$0.0^{10}104$	$0.0^{11}523$	$0.0^{11}260$
7	$0.0^{11}128$	$0.0^{12}624$	$0.0^{12}301$	$0.0^{12}144$	$0.0^{13}682$	$0.0^{13}320$	$0.0^{13}149$	$0.0^{14}688$	$0.0^{14}311$	$0.0^{14}133$

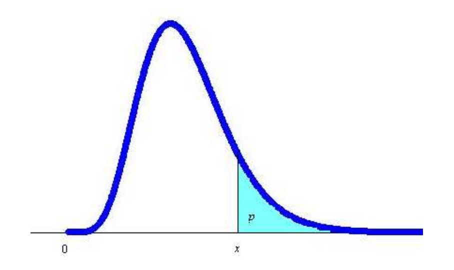
Tabla de t de Student con n grados de libertad



$P\{T > t \mid T \sim t_n\}$

	p											
n	**0.005**	**0.01**	**0.025**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**
1	63.6567	31.8205	12.7062	6.3138	3.0777	1.9626	1.3764	1.0000	0.7265	0.5095	0.3249	0.1584
2	9.9248	6.9646	4.3027	2.9200	1.8856	1.3862	1.0607	0.8165	0.6172	0.4447	0.2887	0.1421
3	5.8409	4.5407	3.1824	2.3534	1.6377	1.2498	0.9785	0.7649	0.5844	0.4242	0.2767	0.1366
4	4.6041	3.7469	2.7764	2.1318	1.5332	1.1896	0.9410	0.7407	0.5686	0.4142	0.2707	0.1338
5	4.0321	3.3649	2.5706	2.0150	1.4759	1.1558	0.9195	0.7267	0.5594	0.4082	0.2672	0.1322
6	3.7074	3.1427	2.4469	1.9432	1.4398	1.1342	0.9057	0.7176	0.5534	0.4043	0.2648	0.1311
7	3.4995	2.9980	2.3646	1.8946	1.4149	1.1192	0.8960	0.7111	0.5491	0.4015	0.2632	0.1303
8	3.3554	2.8965	2.3060	1.8595	1.3968	1.1081	0.8889	0.7064	0.5459	0.3995	0.2619	0.1297
9	3.2498	2.8214	2.2622	1.8331	1.3830	1.0997	0.8834	0.7027	0.5435	0.3979	0.2610	0.1293
10	3.1693	2.7638	2.2281	1.8125	1.3722	1.0931	0.8791	0.6998	0.5415	0.3966	0.2602	0.1289
11	3.1058	2.7181	2.2010	1.7959	1.3634	1.0877	0.8755	0.6974	0.5399	0.3956	0.2596	0.1286
12	3.0545	2.6810	2.1788	1.7823	1.3562	1.0832	0.8726	0.6955	0.5386	0.3947	0.2590	0.1283
13	3.0123	2.6503	2.1604	1.7709	1.3502	1.0795	0.8702	0.6938	0.5375	0.3940	0.2586	0.1281
14	2.9768	2.6245	2.1448	1.7613	1.3450	1.0763	0.8681	0.6924	0.5366	0.3933	0.2582	0.1280
15	2.9467	2.6025	2.1314	1.7531	1.3406	1.0735	0.8662	0.6912	0.5357	0.3928	0.2579	0.1278
16	2.9208	2.5835	2.1199	1.7459	1.3368	1.0711	0.8647	0.6901	0.5350	0.3923	0.2576	0.1277
17	2.8982	2.5669	2.1098	1.7396	1.3334	1.0690	0.8633	0.6892	0.5344	0.3919	0.2573	0.1276
18	2.8784	2.5524	2.1009	1.7341	1.3304	1.0672	0.8620	0.6884	0.5338	0.3915	0.2571	0.1274
19	2.8609	2.5395	2.0930	1.7291	1.3277	1.0655	0.8610	0.6876	0.5333	0.3912	0.2569	0.1274
20	2.8453	2.5280	2.0860	1.7247	1.3253	1.0640	0.8600	0.6870	0.5329	0.3909	0.2567	0.1273
21	2.8314	2.5176	2.0796	1.7207	1.3232	1.0627	0.8591	0.6864	0.5325	0.3906	0.2566	0.1272
22	2.8188	2.5083	2.0739	1.7171	1.3212	1.0614	0.8583	0.6858	0.5321	0.3904	0.2564	0.1271
23	2.8073	2.4999	2.0687	1.7139	1.3195	1.0603	0.8575	0.6853	0.5317	0.3902	0.2563	0.1271
24	2.7969	2.4922	2.0639	1.7109	1.3178	1.0593	0.8569	0.6848	0.5314	0.3900	0.2562	0.1270
25	2.7874	2.4851	2.0595	1.7081	1.3163	1.0584	0.8562	0.6844	0.5312	0.3898	0.2561	0.1269
26	2.7787	2.4786	2.0555	1.7056	1.3150	1.0575	0.8557	0.6840	0.5309	0.3896	0.2560	0.1269
27	2.7707	2.4727	2.0518	1.7033	1.3137	1.0567	0.8551	0.6837	0.5306	0.3894	0.2559	0.1268
28	2.7633	2.4671	2.0484	1.7011	1.3125	1.0560	0.8546	0.6834	0.5304	0.3893	0.2558	0.1268
29	2.7564	2.4620	2.0452	1.6991	1.3114	1.0553	0.8542	0.6830	0.5302	0.3892	0.2557	0.1268
30	2.7500	2.4573	2.0423	1.6973	1.3104	1.0547	0.8538	0.6828	0.5300	0.3890	0.2556	0.1267
31	2.7440	2.4528	2.0395	1.6955	1.3095	1.0541	0.8534	0.6825	0.5298	0.3889	0.2555	0.1267
32	2.7385	2.4487	2.0369	1.6939	1.3086	1.0535	0.8530	0.6822	0.5297	0.3888	0.2555	0.1267
33	2.7333	2.4448	2.0345	1.6924	1.3077	1.0530	0.8526	0.6820	0.5295	0.3887	0.2554	0.1266
34	2.7284	2.4411	2.0322	1.6909	1.3070	1.0525	0.8523	0.6818	0.5294	0.3886	0.2553	0.1266
35	2.7238	2.4377	2.0301	1.6896	1.3062	1.0520	0.8520	0.6816	0.5292	0.3885	0.2553	0.1266
36	2.7195	2.4345	2.0281	1.6883	1.3055	1.0516	0.8517	0.6814	0.5291	0.3884	0.2552	0.1266
37	2.7154	2.4314	2.0262	1.6871	1.3049	1.0512	0.8514	0.6812	0.5289	0.3883	0.2552	0.1265
38	2.7116	2.4286	2.0244	1.6860	1.3042	1.0508	0.8512	0.6810	0.5288	0.3882	0.2551	0.1265
39	2.7079	2.4258	2.0227	1.6849	1.3036	1.0504	0.8509	0.6808	0.5287	0.3882	0.2551	0.1265
40	2.7045	2.4233	2.0211	1.6839	1.3031	1.0500	0.8507	0.6807	0.5286	0.3881	0.2550	0.1265
45	2.6896	2.4121	2.0141	1.6794	1.3006	1.0485	0.8497	0.6800	0.5281	0.3878	0.2549	0.1264
50	2.6778	2.4033	2.0086	1.6759	1.2987	1.0473	0.8489	0.6794	0.5278	0.3875	0.2547	0.1263
55	2.6682	2.3961	2.0040	1.6730	1.2971	1.0463	0.8482	0.6790	0.5275	0.3873	0.2546	0.1262
60	2.6603	2.3901	2.0003	1.6706	1.2958	1.0455	0.8477	0.6786	0.5272	0.3872	0.2545	0.1262
65	2.6536	2.3851	1.9971	1.6686	1.2947	1.0448	0.8472	0.6783	0.5270	0.3870	0.2544	0.1262
70	2.6479	2.3808	1.9944	1.6669	1.2938	1.0442	0.8468	0.6780	0.5268	0.3869	0.2543	0.1261
75	2.6430	2.3771	1.9921	1.6654	1.2929	1.0436	0.8464	0.6778	0.5266	0.3868	0.2542	0.1261
80	2.6387	2.3739	1.9901	1.6641	1.2922	1.0432	0.8461	0.6776	0.5265	0.3867	0.2542	0.1261
85	2.6349	2.3710	1.9883	1.6630	1.2916	1.0428	0.8459	0.6774	0.5264	0.3866	0.2541	0.1260
90	2.6316	2.3685	1.9867	1.6620	1.2910	1.0424	0.8456	0.6772	0.5263	0.3866	0.2541	0.1260
95	2.6286	2.3662	1.9853	1.6611	1.2905	1.0421	0.8454	0.6771	0.5262	0.3865	0.2541	0.1260
100	2.6259	2.3642	1.9840	1.6602	1.2901	1.0418	0.8452	0.6770	0.5261	0.3864	0.2540	0.1260
125	2.6157	2.3565	1.9791	1.6571	1.2884	1.0408	0.8445	0.6765	0.5257	0.3862	0.2539	0.1259
150	2.6090	2.3515	1.9759	1.6551	1.2872	1.0400	0.8440	0.6761	0.5255	0.3861	0.2538	0.1259
200	2.6006	2.3451	1.9719	1.6525	1.2858	1.0391	0.8434	0.6757	0.5252	0.3859	0.2537	0.1258
300	2.5923	2.3388	1.9679	1.6499	1.2844	1.0382	0.8428	0.6753	0.5250	0.3857	0.2536	0.1258
∞	2.5758	2.3263	1.9600	1.6449	1.2816	1.0364	0.8416	0.6745	0.5244	0.3853	0.2533	0.1257

Tabla de Ji cuadrado con n grados de libertad



$P\{X > x \mid X \sim \chi^2_n\}$

							p						
n	**0.01**	**0.025**	**0.05**	**0.10**	**0.15**	**0.25**	**0.5**	**0.75**	**0.85**	**0.9**	**0.95**	**0.975**	**0.99**
1	6.635	5.024	3.841	2.706	2.072	1.323	0.455	0.102	0.036	0.016	0.003932	0.000982	0.000157
2	9.210	7.378	5.991	4.605	3.794	2.773	1.386	0.575	0.325	0.211	0.103	0.051	0.020
3	11.345	9.348	7.815	6.251	5.317	4.108	2.366	1.213	0.798	0.584	0.352	0.216	0.115
4	13.277	11.143	9.488	7.779	6.745	5.385	3.357	1.923	1.366	1.064	0.711	0.484	0.297
5	15.086	12.833	11.070	9.236	8.115	6.626	4.351	2.675	1.994	1.610	1.145	0.831	0.554
6	16.812	14.449	12.592	10.645	9.446	7.841	5.348	3.455	2.661	2.204	1.635	1.237	0.872
7	18.475	16.013	14.067	12.017	10.748	9.037	6.346	4.255	3.358	2.833	2.167	1.690	1.239
8	20.090	17.535	15.507	13.362	12.027	10.219	7.344	5.071	4.078	3.490	2.733	2.180	1.646
9	21.666	19.023	16.919	14.684	13.288	11.389	8.343	5.899	4.817	4.168	3.325	2.700	2.088
10	23.209	20.483	18.307	15.987	14.534	12.549	9.342	6.737	5.570	4.865	3.940	3.247	2.558
11	24.725	21.920	19.675	17.275	15.767	13.701	10.341	7.584	6.336	5.578	4.575	3.816	3.053
12	26.217	23.337	21.026	18.549	16.989	14.845	11.340	8.438	7.114	6.304	5.226	4.404	3.571
13	27.688	24.736	22.362	19.812	18.202	15.984	12.340	9.299	7.901	7.042	5.892	5.009	4.107
14	29.141	26.119	23.685	21.064	19.406	17.117	13.339	10.165	8.696	7.790	6.571	5.629	4.660
15	30.578	27.488	24.996	22.307	20.603	18.245	14.339	11.037	9.499	8.547	7.261	6.262	5.229
16	32.000	28.845	26.296	23.542	21.793	19.369	15.338	11.912	10.309	9.312	7.962	6.908	5.812
17	33.409	30.191	27.587	24.769	22.977	20.489	16.338	12.792	11.125	10.085	8.672	7.564	6.408
18	34.805	31.526	28.869	25.989	24.155	21.605	17.338	13.675	11.946	10.865	9.390	8.231	7.015
19	36.191	32.852	30.144	27.204	25.329	22.718	18.338	14.562	12.773	11.651	10.117	8.907	7.633
20	37.566	34.170	31.410	28.412	26.498	23.828	19.337	15.452	13.604	12.443	10.851	9.591	8.260
21	38.932	35.479	32.671	29.615	27.662	24.935	20.337	16.344	14.439	13.240	11.591	10.283	8.897
22	40.289	36.781	33.924	30.813	28.822	26.039	21.337	17.240	15.279	14.041	12.338	10.982	9.542
23	41.638	38.076	35.172	32.007	29.979	27.141	22.337	18.137	16.122	14.848	13.091	11.689	10.196
24	42.980	39.364	36.415	33.196	31.132	28.241	23.337	19.037	16.969	15.659	13.848	12.401	10.856
25	44.314	40.646	37.652	34.382	32.282	29.339	24.337	19.939	17.818	16.473	14.611	13.120	11.524
26	45.642	41.923	38.885	35.563	33.429	30.435	25.336	20.843	18.671	17.292	15.379	13.844	12.198
27	46.963	43.195	40.113	36.741	34.574	31.528	26.336	21.749	19.527	18.114	16.151	14.573	12.879
28	48.278	44.461	41.337	37.916	35.715	32.620	27.336	22.657	20.386	18.939	16.928	15.308	13.565
29	49.588	45.722	42.557	39.087	36.854	33.711	28.336	23.567	21.247	19.768	17.708	16.047	14.256
30	50.892	46.979	43.773	40.256	37.990	34.800	29.336	24.478	22.110	20.599	18.493	16.791	14.953
31	52.191	48.232	44.985	41.422	39.124	35.887	30.336	25.390	22.976	21.434	19.281	17.539	15.655
32	53.486	49.480	46.194	42.585	40.256	36.973	31.336	26.304	23.844	22.271	20.072	18.291	16.362
33	54.776	50.725	47.400	43.745	41.386	38.058	32.336	27.219	24.714	23.110	20.867	19.047	17.074
34	56.061	51.966	48.602	44.903	42.514	39.141	33.336	28.136	25.586	23.952	21.664	19.806	17.789
35	57.342	53.203	49.802	46.059	43.640	40.223	34.336	29.054	26.460	24.797	22.465	20.569	18.509
36	58.619	54.437	50.998	47.212	44.764	41.304	35.336	29.973	27.336	25.643	23.269	21.336	19.233
37	59.893	55.668	52.192	48.363	45.886	42.383	36.336	30.893	28.214	26.492	24.075	22.106	19.960
38	61.162	56.896	53.384	49.513	47.007	43.462	37.335	31.815	29.093	27.343	24.884	22.878	20.691
39	62.428	58.120	54.572	50.660	48.126	44.539	38.335	32.737	29.974	28.196	25.695	23.654	21.426
40	63.691	59.342	55.758	51.805	49.244	45.616	39.335	33.660	30.856	29.051	26.509	24.433	22.164
41	64.950	60.561	56.942	52.949	50.360	46.692	40.335	34.585	31.740	29.907	27.326	25.215	22.906
42	66.206	61.777	58.124	54.090	51.475	47.766	41.335	35.510	32.626	30.765	28.144	25.999	23.650
43	67.459	62.990	59.304	55.230	52.588	48.840	42.335	36.436	33.512	31.625	28.965	26.785	24.398
44	68.710	64.201	60.481	56.369	53.700	49.913	43.335	37.363	34.400	32.487	29.787	27.575	25.148
45	69.957	65.410	61.656	57.505	54.810	50.985	44.335	38.291	35.290	33.350	30.612	28.366	25.901
50	76.154	71.420	67.505	63.167	60.346	56.334	49.335	42.942	39.754	37.689	34.764	32.357	29.707
55	82.292	77.380	73.311	68.796	65.855	61.665	54.335	47.610	44.245	42.060	38.958	36.398	33.570
60	88.379	83.298	79.082	74.397	71.341	66.981	59.335	52.294	48.759	46.459	43.188	40.482	37.485
65	94.422	89.177	84.821	79.973	76.807	72.285	64.335	56.990	53.293	50.883	47.450	44.603	41.444
70	100.425	95.023	90.531	85.527	82.255	77.577	69.334	61.698	57.844	55.329	51.739	48.758	45.442
75	106.393	100.839	96.217	91.061	87.688	82.858	74.334	66.417	62.412	59.795	56.054	52.942	49.475
80	112.329	106.629	101.879	96.578	93.106	88.130	79.334	71.145	66.994	64.278	60.391	57.153	53.540
85	118.236	112.393	107.522	102.079	98.511	93.394	84.334	75.881	71.589	68.777	64.749	61.389	57.634
90	124.116	118.136	113.145	107.565	103.904	98.650	89.334	80.625	76.195	73.291	69.126	65.647	61.754
95	129.973	123.858	118.752	113.038	109.286	103.899	94.334	85.376	80.813	77.818	73.520	69.925	65.898
100	135.807	129.561	124.342	118.498	114.659	109.141	99.334	90.133	85.441	82.358	77.929	74.222	70.065



Servicio de Publicaciones
Biblioteca Universitaria
C/ Piscinas, s/n
26006 Logroño (La Rioja)
Teléfono: 941 299 187

http://publicaciones.unirioja.es
www.unirioja.es